

T.C.
İSTANBUL 29 MAYIS ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
MÜTERCİM VE TERCÜMANLIK ANABİLİM DALI

LEARNING TO LEARN:
TRANSLATORS' POST-EDITING PATTERNS AND PRIMING

DOKTORA TEZİ

Ayşe Betül SAYIN

Danışman:

Prof. Dr. Işın ÖNER

İSTANBUL

2025

T. C.
İSTANBUL 29 MAYIS ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
MÜTERCİM VE TERCÜMANLIK ANABİLİM DALI

LEARNING TO LEARN:
TRANSLATORS' POST-EDITING PATTERNS AND
PRIMING

DOKTORA TEZİ

Ayşe Betül SAYIN

Danışman:

Prof. Dr. Işın ÖNER

İSTANBUL

2025

TEZ ONAY SAYFASI

T. C.
İSTANBUL 29 MAYIS ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Mütercim Tercümanlık Anabilim Dalı, Çeviribilim (İngilizce) Doktora Programı'nda 010519DR01 numaralı Ayşe Betül SAYIN 'ın hazırladığı “*Learning to Learn: Translators' Post-Editing Patterns and Priming*” konulu doktora tezi ile ilgili tez savunma sınavı, 06/05/2025 günü (11:30 – 13:30) saatleri arasında yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin başarılı olduğuna oy birliği ile karar verilmiştir.

Prof. Dr. Işın ÖNER
İstanbul 29 Mayıs Üniversitesi
(Tez Danışmanı ve Sınav Komisyonu Başkanı)

Prof. Dr. Ayşe Banu KARADAĞ
Yıldız Teknik Üniversitesi

Doç. Dr. Nilüfer ALİMEN
Yıldız Teknik Üniversitesi

Dr. Öğr. Üyesi Sinan Okan ÇAVUŞ
İstanbul 29 Mayıs Üniversitesi

Dr. Öğr. Üyesi Hülya BOY AKYOL
Marmara Üniversitesi

BEYAN

Bu tezin yazılmasında bilimsel ahlak kurallarına uyulduğunu, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, tezin herhangi bir kısmının bu üniversite veya başka bir üniversitedeki başka bir tez çalışması olarak sunulmadığını beyan ederim.

Ayşe Betül SAYIN

06.05. 2025

ABSTRACT

This study aims to measure the priming effect in the evaluation of two neural machine translation outputs. To this end, an empirical experimental setup based on task-based learning was designed. In the first part, theoretical studies on translation competences were examined in detail and the degree to the theoretical work on competences matched the labor market expectations is emphasized. In the second part, quality assessments, which are the textual projections of these competences, are analyzed. In the third part, professional and domain experience affecting the quality assessment process, different expectations in light and full post-editing processes and competences specified by international standards are examined. The fourth part explains the priming effect observed in the context of MT quality evaluation, post-editing and L2 learning behavior. The fifth part explains experimental setup including a pre-experiment questionnaire and the statistical results of the quality evaluation experiment. The results indicate that several hypotheses are confirmed. However, statistical verification of the two hypotheses requires further data. The experimental results show that the priming effect was partially observed. Significant differences were found between the experienced group and the less experienced group in the quality evaluation process. Based on these results, it is stated that future studies can be conducted on other hypotheses that can shed light on theoretical studies on competencies.

Key Words:

translation competences, machine translation evaluation, post-editing, neural machine translation, priming effect

ÖZ

Bu çalışma iki nöral makine çevirisi çıktısının değerlendirmesinde ön hazırlama etkisini ölçmeyi amaçlamaktadır. Bu nedenle, göreve dayalı öğrenme temelli ampirik bir deney düzeneği tasarlanmıştır. İlk bölümde çeviri yetkinlikleri detaylı bir şekilde irdelenmiş ve yetkinliklerin iş piyasası beklentileriyle ne derece örtüştüğü üzerinde durulmuştur. İkinci bölümde bu yetkinliklerin metinsel izdüşümü olan kalite değerlendirmelerine yer verilmiştir. Üçüncü bölümde kalite değerlendirmesi sürecini etkileyen mesleki ve özel alan deneyimi, makine çevirisi sonrası kısmi ve tam düzeltme süreçlerindeki farklı beklentiler, uluslararası standartların belirttiği yetkinlikler irdelenmiştir. Dördüncü bölümde makine çevirisi kalite değerlendirmesi, makine çevirisi çıktısı düzeltme ve ikinci dil öğrenimi davranışları bağlamında ön-hazırlama etkisi tartışılmıştır. Beşinci bölümde deney öncesi anket ve kalite değerlendirmesi deneyinin istatistiki sonuçları açıklanmıştır. Deney sonuçları hipotezlerin bir kısmını doğrulamıştır. İki hipotezin istatistiksel olarak doğrulanması ise daha fazla veri gerektirmektedir. Deney sonuçları ön-hazırlama etkisinin kısmen gözlemlendiğini göstermektedir. Bu gözlemlerde deneyimli grup ve az deneyimli grup arasında belirgin farklılıklar bulunmuştur. Elde edilen bu sonuçlara dayanarak yetkinliklere dair kuramsal çalışmalara ışık tutabilecek başka hipotezler üzerinde ileride çalışma yapılabileceği ifade edilmektedir.

Anahtar Kelimeler:

çeviri yetkinlikleri, makine çevirisi değerlendirmesi, makine çevirisi sonrası düzeltme, nöral makine çevirisi, ön-hazırlama etkisi

ACKNOWLEDGEMENTS

It is difficult to express adequately my appreciation of the help and support offered me by my instructors, my friends, several school authorities and students in the realization of this thesis.

I am most deeply grateful to my thesis advisor Prof. Dr. Işın Öner, who led the study with care expertise and wisdom. She has always been ready to help and full of resources. Without her encouragement, tolerance and understanding attitude, it would not be possible for me to do further research and this study would never have to the shape it has today. I am indebted also to her works which have always been a guide and an inexhaustible source of inspiration.

I wish to express my sincere thanks to Prof. Dr. Ayşe Banu Karadağ for unwavering support even in dark times. Her commitment to science is unshakable and even a global pandemic was not powerful enough to dampen her enthusiasm.

I would like to thank Do. Dr. Nilüfer Alimen for her invaluable help for leading the way to the final presentation with such humor.

I am greatly indebted to Kantan AI for giving me permission to use the interface in my research.

Special thanks are due to my friends, Aslı Kalem Bakkal, Kerem Geçmen, Öznur Yanar Torbalı, Gözde Büklüm and Alper Güneş whose contributions to this thesis are appreciated and must certainly be acknowledged. I also owe special thanks to my family who stand by me always and particularly, dad for giving the invaluable advice acquired for years working globally.

Finally, I wish to express my thanks to the translators and students who participated in the research.

TABLE OF CONTENTS

TEZ ONAY SAYFASI.....	ii
BEYAN.....	iii
ABSTRACT	iv
ÖZ	v
ACKNOWLEDGEMENTS.....	vi
ABBREVIATIONS	ix
LIST OF TABLES.....	x
LIST OF FIGURES	xi
CHAPTER 1 INTRODUCTION.....	1
CHAPTER 2 STRATEGIC COMPETENCE IN EDITING AND REVISION FROM A SELF-LEARNING PERSPECTIVE: LITERATURE REVIEW	9
2.1 Competences and strategic competence in editing	13
2.2 Editing and revising in ISO17100 and post-editing competence in ISO 18587	23
2.3 “Human Translator Competence”: Describing translator’s role in-the loop and the position of revision	27
2.4 Do the quality expectations of the industry correlate with students’ expectations?	46
2.5 Overview: Do social constructivism and stone-hard truths of the labor market align? or human at the core vs. human in the loop	51
CHAPTER 3 PREFERENTIAL CHANGES AND EDITING THRESHOLD: EXPERTISE IN DOMAIN AND THE NOVICE TRANSLATOR	53
3.1 From “optimal equivalence” to MQM.....	54
3.2 Requirements for full and light post-editing according to ISO 18587	68
3.3 Light post-editing and full post-editing: The discrepancy between ISO in terms of definitions and what different experience groups deem necessary for each level of post-editing	70
3.4 Preferential changes and inter-annotator agreement.....	104
3.5 Domain and PE effort: Evaluation of NAATI specialized texts and “non- specialised domains”.....	111
3.6 Sentence-level evaluation or document-level evaluation?.....	113
3.7 Quality over quantity?	117
CHAPTER 4 THE PRIMING EFFECT AND EXPERTISE	123
4.1 What is the “priming effect”?	124
4.2 Priming effect and L2 learning	126
4.3 Priming effect and MT.....	128
CHAPTER 5 A 2-TIERED EXPERIMENT ON PRIMING WITH SPECIALIZED TEXTS	143
5.1 Sustainability reports and their stylometry	143
5.2 MT systems used and preparing the priming set	145
5.3 The survey results	150
5.4 Iterative text revision and editing	157
CHAPTER 6 CONCLUSION.....	193

REFERENCES	196
ANNEX I: THE TRANSLATION BRIEF	223
ANNEX II: ST AND TRANSLATIONS USED FOR THE EXPERIMENT	225
ÖZGEÇMİŞ	232



ABBREVIATIONS

Abbreviation	Meaning
DGT	The Directorate-General for Translation
DQF	Dynamic Quality Framework
ELIA	European Language Industry Association
EMT	European Master's in Translation
EUATC	European Union of Associations of Translation Companies
FIT	International Federation of Translators
GALA	The Globalization and Localization Association
GRI	Global Reporting Initiative
LIND	The Language Industry Expert Group
LISA	Localization Industry Standards Association
LLM	Large Language Models
NAATI	National Accreditation Authority for Translators and Interpreters
NMT	Neural Machine Translation
PBMT	Phrase-Based Machine Translation
RBMT	Rule-Based Machine Translation
SMT	Statistical Machine Translation
TAUS	Translation Automation User Society
UCT	User-centered Translation

LIST OF TABLES

Table 1:A List of Revision and Post-Editing Competences	22
Table 2: MQM Full Error Categories updated in 2024	74
Table 3: Automated Scores Distribution for Kantan Pretrained Engine	146
Table 4: Examples from Priming Segments	147
Table 5: Sketch Engine Analysis on ST, MT-Priming Set and MT-Test Set.....	149
Table 6: Descriptive Statistics for All Participants.....	164
Table 7: Error Issue Type Annotations.....	178
Table 8:Test of Fixed Effects.....	179
Table 9: Estimates of Fixed Effects	181
Table 10: Descriptive Statistics of Total Quality, Time Spent on PE and PE Effort	183
Table 11: Paired Samples Test.....	184
Table 12: ANOVA Test Results	185
Table 13: Calculations for Cohen's D.....	186
Table 14:Time Spent on PE by Professionals.....	187
Table 15: Time Spent on PE by the Less Experienced Group.....	188

LIST OF FIGURES

Figure 1: Revision Competence Model by Scocchera.....	17
Figure 2: A Comparative Study of Light PE Guidelines presented by Hu and Cadwell (2016).....	92
Figure 3: A Comparative Study of Full PE Guidelines presented by Hu and Cadwell (2016)	94
Figure 4: Answers given for Q2	151
Figure 5: Answers given for Q10	155
Figure 6: Answers given for Q11	156
Figure 7: Working Interface in Matecat for the Participants	159
Figure 8: A Revised Segment by Participant 2.....	189
Figure 9: A Revised Segment by Participant 4.....	190
Figure 10: A Revised Segment by Participant 5.....	190
Figure 11: A Revised Segment by Participant 5.....	191
Figure 12: The Same Segment Annotated in the Test Set	191
Figure 13: Revising the Passive with an Active Voice.....	191

CHAPTER 1 INTRODUCTION

*Every algebraic equation is an icon, insofar as it exhibits
by means of the algebraic signs (which are not themselves icons)
the relations of the quantities concerned¹.*

Translation from scratch, post-editing, monolingual paraphrasing, summarization, machine translation output through pivot language, search engine optimization, copywriting, automated production of technical documentation, backtranslation, copyediting, etc. are all components of the process of present-day content production. Accompanied by the exponential growth in data-processing technologies, advancements in the language industry have yielded an unprecedented transformation for the linguist/translator in the last decade. Until recently, the natural language processing technologies such as data retrieval and data augmentation were barely visible to the translator/interpreter and content writer merely through plagiarism and/or language quality assurance (LQA) tools, whereas with today's standards even the freelance translator working in the most 'remote' environment would not be able to avoid the discussions about machine translation and use of data. In this respect, two questions arise, one for the language industry marketplace and the other for translation pedagogy; how we should evaluate the outcomes of this developing technology in terms of pricing, quality and processes of content creation, and how this constantly developing technology complements current Translation Studies (TS) curricula.

This thesis sets out to provide answers for both questions with an empirical two-tiered experiment on the post-editing of a standardized technical text with both novice and experienced translators on the first tier, and the result of the priming process on both groups on the second. In order to conduct this two-tiered experiment, though there are topics that need to be discussed to meticulously analyze the research results. Starting with the hypothesis behind this experiment, the literature review in Chapter 1 provides the background information for understanding what various designations in the translation/localization

¹ C.S. Pierce quoted from Roman Jakobson's article dated 1965 titled "Quest for the Essence of Language", *Diogenes* 13, no. 51 (1965): 21-37, <https://doi.org/10.1177/039219216501305103>.

workflow are used in the current international standards and white papers. In line with the current EMT Competence Framework published in October 2022, as it “acknowledges that machine translation (MT) represents a growing part of translation workflows”, it is essential for any curriculum to consider “MT literacy and awareness of MT’s possibilities and limitations is an integral part of professional translation competence” (EMT 2022,7). Therefore, rather than simply analyze the effect of experience on post-editing quality, this study aims to observe and report the analytical results of learning from others’ post-editing, hence using MT output as a pivot for teaching editing and revising. By combining two skills, i.e., skills 11 and 14 of ‘Translation’ competence, below iteratively, this study attempts to provide further research results, the previous attempts at which have remained “contradictive” to this day (Konttinen, Salmi and Koponen 2021,194-95). Before analyzing the research results though, we need to understand these skills mentioned above. These two skills are:

11 Check, review, revise and evaluate their own work and that of others according to standard or work-specific quality objectives and assess the appropriateness of using tools for the work at hand

14 Post-edit MT output using style guides and terminology glossaries to maintain quality standards in MT-enhanced translation projects (EMT 2022, 8)²

Though seemingly holistic and complementary, these two skills have not been studied in relation to one another in machine translation research. Afterall, the skillset a translator should acquire according to “ISO 17100 Translation Services” presents major differences when compared to the skillset “ISO 18587 Translation services — Post-editing of machine translation output — Requirements”. The fundamental difference between the two standards is stated in the first one, being “The use of raw output from machine translation plus post-editing³ is outside the scope of this [ISO 17100] International Standard” (ISO 17100:2015, 1). So, by combining the observable results of the experiment on the acquisition of these two

² The order of these skills will be discussed below in Chapter 2.

³ ISO 18587 defines “post-edit” as “edit and correct machine translation output” (2017, 2).

skills by novice translators this thesis attempts to shed light on two aspects of the learning in the translation process, i.e, revising⁴ your own work and post-editing. But before delving deeper into the reciprocal correlation of these two skills, we need to examine the ‘contradictory’ research results obtained in previous studies. In view of the literature review conducted in Chapter 2, the subject matter in discussion here can be put into words as the “priming effect” of post-editing on editing and revision quality. Used in psycholinguistics and studies on bilingualism, what exactly is priming? The editors of one of the most recent studies on the subject matter, Trofimovich and McDonough, define “priming” as “the phenomenon in which prior exposure to specific language forms or meanings either facilitates or interferes with a speaker’s subsequent language comprehension or production” (2011, 3). Following this point, it can be comfortably assumed that exposure to raw machine output ‘primes’ novice translators’ understanding of translation quality to a certain degree, regardless of the fact that this might constitute positive or negative priming from the professional translators’ point of view.

In order to understand the hypotheses of the present study, research results on priming in second language (L2) learning can be compared to differences in machine translation evaluation by two groups of translators, with a range of professional experience levels from complete novice to seasoned professionals. Yet, this evaluation concerns varying levels of competence. Though there is no such classification in the current literature⁵, the linguistic layers⁶ of evaluation metrics such as Multidimensional Quality Metrics and TAUS Dynamic Quality Framework can be easily interpreted along the lines of research on priming in syntax and lexicon. Therefore, putting aside earlier uses of MT systems “mainly for gisting purposes”, the present study argues for a panoramic approach to studying translation technology in use by looking both at L2 teaching and teaching editing that is directly linked to post-editing (Resende, Cowan and Way 2020). In order to understand varying uses of MT

⁴ ISO 17100 defines “revision” as “bilingual examination of target language content against any source language content for its suitability for the agreed purpose” (2015, 2).

⁵ It can be argued that translation quality evaluation metrics are in line with PACTE’s language and extra-linguistic sub-competence categories (2003).

⁶ Here I deliberately use unique terminology.

systems, we hence need to analyze them as “a tool supporting writing skills, grammar skills and language production in a second language (L2)” (245). A similar line of thinking has been studied in the literature with the main questions being whether people with varying degrees of linguistic competence can be post-editors and what sort of competence is “adequate” for post-editing and revision. In this respect, these studies question the assumptions about the difficulty of each task in any translation project, including TEP⁷, which also opens the way for AI-based assignment of each task without the services provided by the project manager. For grading the difficulty of tasks between “language content comprehension” to “language content production”, studies in the current literature use different techniques asking similar questions to the above mentioned, and ironically enough, their results contradict one another. For instance, Kontinen, Salmi and Koponen state in a seminal article:

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 (for example, Way 2009; Mossop 2020), whereas post-editing is sometimes envisioned as a task for beginning translators or even non-translators familiar with the subject matter. For example, Guerberof Arenas (2014b) found no significant differences in the post-editing competences of experienced translators and translation students, and some studies (for example, Zaidan and Callison-Burch 2011; Mitchell 2015) have shown potential for non-translators as post-editors in specific situations. On the other hand, Yamada (2015) found that foreign language students (as opposed to translation students) did not reach overall post-editing quality comparable to professional translators, and others too have argued for post-editing as a specialised task suitable for translators (see O’Brien 2002) (2021, 194-95).

These divergent results from studies employing varying competence groups are due to the fact that earlier studies had technological limitations for measuring technical and temporal effort and were to certain degree biased because of the experiment setup. The details of these earlier studies will be further explored in Chapter 2, but there is only one remaining study on

⁷ TEP is the industry acronym for translation, editing and proofreading (ISO 17100: 2015).

the priming effect which needs to be discussed to understand the hypothesis of the present one.

Resende et al. (2020) applies “‘a syntactic priming experimental paradigm’ to a pretest-priming design which compares ‘participants’ linguistic behaviour before and after a translation task using Google Translate” (1). They define “syntactic priming effect”, or “syntactic alignment” or “structural alignment”, as the incidence when “people, in a communicative context, repeat the same syntactic structure previously seen, heard or read” (2), following J. Kathryn Bock’s pioneering study on the syntactic repetition effect conducted in a laboratory in 1986. This study consisted of two phases, namely a prime phase and a test phase. In the former, participants were asked to read a sentence and repeat the same sentence out loud. After this ‘prime’ phase, participants were asked to describe an image. The assumption of the study was that the participants would produce the same structure they had just repeated in the prime phase in the subsequent utterance. Bock (1986) observed that in the test phase “participants tended to use the same syntactic structure previously heard and repeated”, hence affirming the hypothesis (Resende et al. 2020, 246). However, a number of studies also revealed an inverse preference effect, in which case “less frequent syntactic structures prime more than more frequent structures” (Ferreira and Bock 2006). Similar studies demonstrating syntactic alignment, thus “repetition effect” (Heyselaar et al., 2017a), between humans and computers (Branigan et al., 2010; Cowan et al. 2015) as well as humans and avatars⁸ (Heyselaar et al. 2017b) can be found in the literature. Therefore, in an interactive context it is clear that interlocutors tend to prime on the syntactic level, and hence syntactically align, with each other “both in first and second languages as well as with computer partners” (Resende et al. 2020, 2). But then, this particular sort of priming does not tell us much about semantic alignment, in other words it does not explain the fact that the highest number of errors is found in the category of terminology in almost all studies with

⁸ Heyselaar et al. challenge the assumption that humans do not interact with computers in the same way they interact with other humans by “showing that language processing in interaction with a human-like virtual agent (“avatar”) is comparable to interactions with a human partner” (2017b, 46).

participants including students/novice translators (Guerberof Arenas 2014; Yamada 2014 and 2019). Though a repetition effect can also be observed in the behavior of professionals over a rather longer period of time, the errors in the terminology category in MQM can catch the novice translator's attention, and a professional would be quick to spot it immediately. Following this line of thought, there are studies where participants' attention is tested in a mutually exclusive hypothesis. That is, they either focus on syntax or semantic unknowns if they are insufficiently proficient/competent. In this respect, Resende et al. (2020)'s study is one of the few works using a similar hypothesis, i.e., "syntactic priming experimental paradigm", measuring the comparison between participants' linguistic behavior before and after a translation task using Google Translate. In this study, researchers summarize their point thus:

We hypothesize that the effect of English proficiency observed is related to the participants' focus of attention when processing the second language. A number of studies (e.g. Marinis et al. (2005)) have shown that language learners at lower levels of proficiency are less sensitive to syntax than more proficient bilinguals because they focus their attention more on resolving semantic ambiguities than on resolving parsing problems. However, at higher proficiency levels it is possible to observe more automaticity in second language parsing. Accordingly, our results suggest that MT systems can be useful in English language learning as it can facilitate end-users to access or construct problematic syntactic structures in English due to the structural differences between the languages. (Resende et al. 2020, 252)

Similar to the findings of this research, the present study will attempt to adapt the categories in the MQM framework to linguistically-oriented studies and, therefore, will attempt to avoid ambiguities in the error annotation. Thus, the experiment of this thesis will be conducted as in the following scheme. In tier 1, higher domain knowledge, represented by the higher number of evaluation errors in the terminology category in post-editing and signaled by the variable of experience, which is higher in comparison to the novice group, is hypothesized to have a higher priming effect on the editing decisions taken by the novices. In tier 2, this study attempts to answer which source of information, i.e. primed sentences or raw MT output, will have a higher priming effect for students. Thus, we will be able to learn what

type of exposure is more efficient for which level of competence performed by the novices and professionals. In order to improve the results of this study, further research can be conducted using the results of a contrastive analysis of the priming effect observed in L2 and L3, opening ground for a more ‘sustainable’ use of MT systems and the value of domain-knowledge as the opposite of linguistic competence in the dialectical structure of MT evaluation schemes.

The questions this thesis sets forth to answer are as follows, starting with the hypotheses behind them:

- Post-editing distance of students with higher level of experience in translation will demonstrate a decrease in the number of errors and tend to revise/revert to previous translations; less so with respect to reference translation.
- Do we observe a correlation between students’ perception of MT output and their priming?
- Are they primed by altered Google Translate output? Do we observe priming in their post-editing behavior?
- As the results of the previous studies indicate, if post-editing behavior is more consistent across informants, and the low to moderate correlation between post-editing behavior and human evaluation scores is observed, then *post-editing similarity* might prove to be a more reliable measurement than human evaluation for gaining further insight into MT quality levels when participants are asked to do full post editing (Depraetere et al. 2014, 99).
- The main question of this thesis is: How well are novice translators primed from seemingly high-quality MT output in terms of expected translation quality? Do they understand the capabilities and limitations of MT output, question accuracy of MT systems? For instance, how far do they go in their search for terms used in translation during both the post-editing and the revision process? Moreover, how do they search for terms, and do they have any specific techniques?

In this respect, Chapter 2 focuses on the literature review behind the hypotheses of this thesis. The results of similar priming studies are compared in this chapter, laying bare the

‘contradictions’ in the results and the biases in research most commonly observed in post-editing experiments. Lack of connection to the internet and resources to check whether a terminological unit is correct in a certain context might be a very common bias, hence causing incorrect terminology errors to be highest in number. In Chapter 3, the literature review on preferential changes and editing threshold will be analyzed in detail and the reasons for lack of correspondence between minute levels of analysis will also be discussed. Chapter 4 examines the structure of data in the corpus on sustainability and the importance of the subject matter for the translation industry. Focusing on the rhetoric of sustainability reports as standardized documents and article abstracts, it allows us to compare two discursively-different text categories with the help of language quality assurance tools. It also sheds light on how to control the batch size of a certain corpus. Chapter 5 describes the experimental setup, examines and discusses the search results. The two-tiered experiment provides empirical evidence for teaching translation, editing and revision skills. The final chapter, Chapter 6, attempts to answer the following questions: Is raw MT output an alternative for teaching editing and revision? How effective is the simulation of using MT as a fellow translator? What sort of a correlation can we establish between editing, revision and post-editing skills? Given this empirical evidence, how should translation studies instructors reform their course objectives and syllabi in light of current best practices in the industry for future advancements in translation technology?

CHAPTER 2 STRATEGIC COMPETENCE IN EDITING AND REVISION FROM A SELF-LEARNING PERSPECTIVE: LITERATURE REVIEW

Let's start from the very beginning. Any job or profession by definition and by practice requires people who practice it to have some “competences” and “skills”. This chapter first discusses the ‘basic’ competences of a translator as an individual working in a translation workflow and then focuses on the alternating interpretations of “translator competence” in a dynamic marketplace. The literature review below provides a detailed but concise overview of approaches to what constitutes “strategic” behavior in a professional in a demanding ecosystem. In terms of generic definitions, the European Classification of Skills, Competences, Qualifications and Occupations, (ESCO), applies the same definition of “competence” as the European Qualification Framework (EQF)⁹. According to this framework, competence is defined as “the proven ability to use knowledge, skills and personal, social and/or methodological abilities, in work or study situations and in professional and personal development”. In the translation workflow context, the concrete skills included for various competence categories has been widely discussed over a period of time now with the integration of machine translation in the workflow (see Pyne 1994). Yet, the pedagogical aspect of professional development requires meticulous empirical research data in order to observe changes in the development of abilities. The literature review below demonstrates that such research data was not hitherto available and not distributed until recently in the translation studies discipline. Therefore, the iterative phases in the ISO standard for translation workflow sheds light on what should best be measured for both teaching purposes and the fair integration of post-editing in the workflow.

After the initial phase of drafting either a manuscript or any other type of work involving translation – even designing a user interface or any sort of localization project –, what follows is termed some kind of ‘editing’ and ‘revision’ to a varying degree. According

⁹ “Competence,” European Skills, Competences, Qualifications and Occupations, European Commission, accessed April 26, 2025, <https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/competence>.

to ISO 17100 workflow, it is termed “translate-edit-proofread (TEP)”, which is based on the four-eyes principle with ‘four-eyes’ referring to two people working on the same project with review and proofreading as a non-compulsory phase. At the end of the translation phase, the standard specifies that the ‘translation’ task “shall at least include the translator’s overall self-revision of the target content for possible semantic, grammatical and spelling issues, and for omissions and other errors, as well as ensuring compliance with any relevant translation project specifications” (17100: 2015, 10). Because this study aims to provide empirical data towards post-editing and revision practices of novice and seasoned translators, it is worth noting how it is defined for practice and development in both foreign and native languages as a specific language competence in canonical manuals. In this respect, though not compliant with ISO standards, the *Chicago Manual* states that

Manuscript editing, also called copyediting or line editing, requires attention to every word and mark of punctuation in a manuscript, a thorough knowledge of the style to be followed, and the ability to make quick, logical, and defensible decisions (Chicago 2.49).

These “defendable decisions” are evidently a result of both the context knowledge and linguistic craftiness of an individual. Yet, how this craft is mastered with machine translation in the workflow and the step-by-step decision-making process underlying it needs to be documented in order to understand the new skills translators need to acquire in a world of NLP technologies like large language models. Another reference to editing outside the localization sphere can be found in a preeminent manuscript production professional’s guide. According to Scott Norton, for mostly non-fiction content in the industry a unique term, “developmental editing”, can be used. But it is applicable when “significant structuring or restructuring of a manuscript’s discourse” takes place (2009, 1). Norton very practically explain what this task involves as below:

In the industry, opinions vary as to what constitutes “significant” restructuring. At the University of California Press in Berkeley, we define developmental editing as intervention that moves content from one chapter to another, or rearranges the lion’s

share of a chapter's contents within itself, but that falls short of writing new material (2009, 2).

On this point, it is crucial to note that the *Chicago Manual* differentiates “manuscript editing” from its developmental counterpart as below:

It [Manuscript editing] is distinct from developmental editing (not discussed in this manual), which more directly shapes the content of a work, the way material should be presented, the need for more or less documentation and how it should be handled, and so on (2.48).

In line with this categorization, the *Chicago Manual* differentiates between two types of editing, namely mechanical editing and substantive editing. On the one hand, “mechanical editing involves the consistent application of a particular style¹⁰ to a written work” (2.49) and the correction of errors related to grammar, syntax, and usage. Substantive editing, on the other hand, is about “the organization and presentation of existing content”. In this case, as the name implies, the content needs to be rewritten to be further aligned with the style or ambiguities need to be eliminated, and hence the content is reorganized or tightened to make disorganized or loose sections to fit better with the whole and, if necessary, tables are adjusted or recast (2.50). In a similar vein, an eminent figure on editing, Brian Mossop elaborates on “four broad types of checking/amending work”, namely copy editing, stylistic editing, structural editing, and content editing (2020, 17). Putting aside the structural side, including “reparaphrasing” and content editing, he differentiates between five levels of editing apart from the reading effort. While on level 1 the editor is asked to make no changes, on level 5 the workflow (“the division of labor” in Mossop’s words) requires the editor to rewrite the text (27-28). Between these levels lie “make changes for correctness”, “make changes for intelligibility”, and “make changes for writing quality (‘style’)” respectively. Mossop’s degrees of editing are valuable in terms of the discussion about post-editing, self-revision, revising the work of others that is held below.

¹⁰ Style here refers to “rules related to capitalization, spelling, hyphenation, and abbreviations; punctuation, including ellipsis points, parentheses, and quotation marks; and the way numbers are treated” (Chicago 2.49).

The technical straightforwardness of these naming and competences hides the empirical technical skills required for these processes of content creation. Though in terms of translation revision and review, the tasks involved reflect a more empirical standpoint in the case of specific skills and their application. According to ISO 17100, revision is defined as “bilingual examination of target language content against source language content for its suitability for the agreed purpose”, which can also be called “bilingual editing” and review is defined as “monolingual examination of target language content for its suitability for the agreed purpose”; it can also be called “monolingual editing” (ISO17100: 2015, 2). At this point, it is crucial to note how “post-edit” is differentiated from editing fuzzy matches found in a translation memory. The standard here clearly states that “the post-editor will edit output automatically generated by a machine translation engine. It does not refer to a situation where a translator sees or uses a suggestion from a machine translation engine within a CAT tool” (ibid). Therefore, it is obvious that the standard considers reviewing and proofreading of fuzzy matches in a translation project and post-editing to be comparable but distinct tasks. What is involved in the post-editing task text-wise has been discussed even before the integration of MT systems to CAT tools. As a consequence, the demands of post-editing competences¹¹ have been discussed both didactically, i.e. the international standard and research groups, and empirically, i.e. research on post-editing by scholars and the industry with students and professionals.

This chapter explores first how competences are conceptualized in the Translation Studies literature and how studies tend to overlap and disagree with one another in terms of “post-editing competence” and then delves into the depths of ISO18587 for a detailed analysis of what the international standard demands for translation services related to post-editing. It later brings the concept of “human translator competence” onto the table and asks whether the expectations from the language services industry align with students’ and scholars’ point of view. In the overview section at the end of the chapter, the focal question of this literature review is addressed.

¹¹ ISO 17100 defines ‘competence’ as the “ability to apply knowledge, experience, and skills to achieve intended results” (2015, 4).

2.1 Competences and strategic competence in editing

Theoretical work¹² on “translation competence” is abundant in Translation Studies. Self-observation by someone who is both a translator and researcher and works on a certain translation, and reporting the steps (s)he takes might seem the easy way to go. But when ‘that’ someone teaches these specific steps to other people, it is very straightforward to understand that others do not follow the same pattern. How, then, are we going to set standards, considering that we need them in order to practice a certain profession in a global marketplace? The first approach that comes to mind is to observe a number of people’s working practices and try to report on these empirical findings to others in the same profession in an objective manner. For this reason, research into Think-Aloud-Protocols and post-editing has been gaining momentum since the start of millennium, particularly with publication of PACTE research group’s findings on their re-modelled translation competence test in 2003. Yet, first and foremost, postulates of translation competence need to be explored in order to comprehend how empirical data is theoretically formulated on them. The PACTE model foregrounds the learning process behind competences and states that “translation competence acquisition is a process of restructuring and developing sub-competencies of translation competence” (2003, 8). This acquisition model is defined as a “dynamic, spiral process that evolves from novice knowledge (pre-translation competence) to expert knowledge (translation competence)”, with two other essential elements in which “the development of procedural knowledge and, consequently, of the strategic sub-competence are essential” and “the translation competence sub-competencies are developed and restructured” (ibid). As expected from a model of learning, it comprises four sub-competencies which

¹² Hilal Öztürk Baydere PhD thesis on the transformation of the translation profession in the context of digitalization offers a comprehensive survey of translation competences and analyzes the curricula of translation departments in Türkiye offering undergraduate courses and “National Occupational Standard Translator/Interpreter Level 6” document in terms of entrepreneurial competence (Öztürk Baydere 2022).

(1) [are] inter-related and compensate for each other; (2) do not always develop in parallel; (3) are organised hierarchically; (4) variations occur in relation to translation direction, language combinations, specialisation and the learning context (8).

There are clearly “relations, hierarchies and variations” in the process of acquiring these sub-competences but, very briefly, the PACTE group labels them as bilingual, extra-linguistic, strategic, instrumental, and knowledge about translation sub-competences with psychophysiological components. Starting with the initial step on a micro-scale ‘bilingual sub-competence’ has elements related to “pragmatic, socio-linguistic, textual and lexical-grammatical knowledge in each language”, extra-linguistic sub-competence “encyclopaedic, thematic and bicultural knowledge”, translation knowledge “knowledge of the principles that guide translation (processes, methods and procedures, etc.) and the profession (types of translation briefs, users, etc.)”, instrumental “knowledge related to the use of documentation sources and information technologies” used in the translation process and, overarchingly, strategic sub-competence includes problem-solving skills (2005, 610). Those elements, including memory, attention span and critical thinking skills, are grouped under psychophysiological components.

The PACTE group states that their research does not focus on bilingual and extra-linguistic sub-competences because these two can be possessed by bilinguals who do not necessarily have translation competence (PACTE 2008, 108). One criticism that could be made for the PACTE model is that this general abstract modelling of competence acquisition does not deal with micro-learning behavior among novice translators and, therefore, does not explore editing and revision ‘skills’ in detail. What PACTE tried to achieve was to shed some light on the translation process by conducting an empirical study with internal and external support through a test study with three expert translators and three language teachers (2005, 615). Though the criticism about lacking concrete skills can still be made about PACTE, it can be inferred from the table presented that translation knowledge and strategic sub-competences deal with revision and editing as the indicators of these sub-competences label “stages in the translation process” and “time dedicated to each stage (orientation,

development, revision)” (613). Yet, the test study does not include variables for measuring them.

In contrast, ISO 17100 provides a more concrete categorization of professional competences even though they match one another on a surface level. What ISO 17100 suggests in terms of competence categorization is grouped under six competences, namely translation competence, textual and linguistic competence in source and target languages, competence in research, information acquisition and processing, cultural competence, technical competence and lastly, domain competence (2015, 6). As for revisers, the standard very clearly states that they should have the same competences the translators have (10).

Another problematic issue with PACTE’s postulates can be made about the necessary period of time for learning or acquiring any competence. But the PACTE group addresses this issue in their paper. Regarding the learning process of revision and also other translation processes, they claim that “depending on the language combinations, the process may be more or less rapid, or, depending on the translation speciality (legal, literary translation, etc.) one sub-competence may be more important than another” (2003, 8). Thus, when it comes to post-editing the questions of which sub-competence is more important or open to change in the case of practice, and also, the acquisition of which sub-competence is going to be more rapid or not given the context of the current translation studies pedagogy need to be critically evaluated with empirical studies – which is what this thesis attempts to do.

When reviewing the literature for competences, the change in scope of competences needs to be questioned in the light of the trajectory for post-editing and related skills. Therefore, the question posed by Kalle Konttinen, Leena Salmi, Maarit Koponen becomes increasingly relevant: “Revising a (human) translation and post-editing MT output both require specific skills and competences, but are those competences the same, different or overlapping, or is one perhaps a subset of the other?” (2021, 190). Following this question, they set out with a compare-and-contrast study on five models of competences¹³ that includes

¹³ As mentioned beforehand, the list of competences for translation is far greater in number. Senem Öner Bulut (2019) provides a more comprehensive list of studies analyzing and proposing models for translation and other related services. However, because most of these models were formulated before MT started to be

MT and TM skills in it by Scocchera (2017, 2019), Künzli (2006), Mossop (2020), Rico and Torrejón (2012), and finally Pym (2013). Starting with Robert et al. (2017), which combines the PACTE model and the Göpferich's TransComp, this study highlights the fact that not much research is done on translation revision and drawing from Hansen (2009), stresses that "translation revision seems to require additional skills, abilities and attitudes, and/or enhanced levels of competence in certain areas" (Hansen 2009 cited in Robert et al. 2016, 2). Therefore, because they address the need to build a 'multicomponential' model such as the ones by PACTE and Göpferich, they borrow bilingual and extralinguistic subcompetences and also knowledge-about-translation subcompetence from the PACTE model and translation routine activation and the tools and research subcompetence from Göpferich. At the center of their model lies "strategic subcompetence" with three revision-related other subcompetences. These are Göpferich's "tools and research subcompetence (including psychomotor subcompetence)", "interpersonal subcompetence" and "revision routine activation subcompetence" (Robert et al. 2016, 13). The hypothesis of their research compares the results of an experimental study with those people who took their specifically designed course for revision and those who did not. Thus, underscoring the interpersonal skills necessary for revision, they try to address the need to "achieve the mental switch from a 'retranslating' to a 'revising frame of mind'" (Mossop from Robert et al. 2017, 2). The method of this study is to measure the number of hyperrevisions and hence, assess the revision quality. Strategic subcompetences in this study function in the planning and carrying out of the revision project ("selecting the most adequate procedure in view of the task definition, reading for evaluation, applying a detection strategy [and] applying an immediate solution or problem-solving strategy, making only the necessary changes") and also justifying textual changes. When it comes to Scocchera's meticulous theoretical and empirical study on translation revision and creation of a model for students' acquisition of the practice, her work also presents a multicompetence model of revision which has a partly similar, partly unique set of components to the previous research (2017). This revision

used in TM systems and the international standard for post-editing was published in 2017, the present study will not delve into their detail.

stage, students are asked to review and comment on the revision work done up to that stage and explain their decisions to a hypothetical translator. Scocchera elaborates to the effect that this specific stage has an impact on future revision styles and also, looks at one's own decisions from a critical point of view. Therefore, the hypothesis of this stage in the whole study was to

having to add comments to one's revision work calls for an extra amount of analytical and critical thinking on the reviser's part and this may result in new revision changes even at this third stage of the revision activity; not being provided with any specific training on what to say and how to say it might hinder student revisers from exploiting their full commenting potential, thus limiting the number of comments made (2019, 12).

After students are given theoretical training into error taxonomies and classifications toward the end of the 10-week course, students were asked to comment on their on their revision work again. The purpose of this final stage was to compare the qualitative and quantitative data on revision comments and see whether students' acquisition of revision subcompetences were positively marked in "the overall number of comments, the number of formulations using specific metalanguage, and the number of diplomatic formulations" (14). Apart from these studies with their focus on the psycho-physiological competence summarized above, two other names in the literature foreground the linguistic aspect of revision as almost a middle ground where the decision is taken either to translate from scratch – retranslate – or work on the text with minimum change. The first person in this group is Brian Mossop who presents a practice-oriented set of abilities for revisers. Practices in this set involve detecting problems in a translation, "knowledge of revision procedures", "ability to make decisions quickly", "ability to revise to varying degrees", "ability to justify the need for changes and the related ability to avoid unneeded (unjustifiable) changes", "ability to correct through small changes rather than retranslation", "ability to appreciate other people's approaches", "ability (as a trainer) to diagnose a translator's weaknesses (and strengths!) and give advice on how to improve", with three other personal qualities, "diplomacy to deal with conflict", "leadership to organize a group of translators working on a project", and "cautiousness to

avoid making the translation worse and to keep in mind one's own limitations" (118-119.) at the other end of the spectrum of research focusing on the changing landscape of linguistic skill sets and practices is Anthony Pym. After positing that the EMT multi-componential model of competences is "forever condemned to lag behind both technology and the market" (Pym 2013, 490), his list of 'things' that have the potential to "upset" this model includes his prediction:

If we limit ourselves to that frame, the impact of TM/MT is relatively easy to define (see Pym 2011b): whereas much of the translator's skill-set and effort was previously invested in *identifying* possible solutions to translation problems (i.e., the *generative* side of the cognitive process), the vast majority of those skills and efforts are now invested in *selecting* between available solutions, and then adapting the selected solution to target-side purposes (i.e., the *selective* side of the cognitive processes). The emphasis has shifted from generation to selection (2013, 493).

But this focus from generation to selection brings with it several problems along the blurred lines. First and foremost, don't we need to consider translation different from revision or does post-editing replace revision? Then, the other major problem is related to the experience of both translator and the reviser. Pym's lists 10 skills needed in the "Machine Translation age", which combines MT and TM skills" (Kontinen et al. 2021, 191) under three larger categories of skills. These larger categories are "learn to learn", "learn to trust and mistrust data" and "learn to revise translations as texts" (Pym 2013). In addition, the skills and competences the translator is expected to have focus on target language to a great extent (as opposed to area or domain knowledge other researchers underscore) as much as interpersonal communicative skills. Ironically enough, empirical studies on different levels of post-editing are going to demonstrate totally contrasting results in some cases in the next chapter, but Pym's discussion is still worth noting as he states

Area knowledge ("thematic competence" [EMT Expert Group 2009]) should be affected by this same logic. Since TM/MT reduces the need for language skills, or can make the need highly asymmetrical, much basic postediting can theoretically be done by area experts who have quite limited foreign-language

competence. This means that the language expert, the person we are still calling a translator, could come in and clean up the postediting done by the area expert. That person, the translator, no longer needs to know everything about everything. What they need is great target-language skills and highly developed teamwork skills (2013, 491).

So when the translator is expected to “clean up” after an area expert post-edits a text, expert replaces translator and the self-revision the international standard expects from the translator is either covered by the expert or the translator-cum-reviser. The translation step in the production process is minimized into small changes. But on the contrary, what if no area expert gets involved in the process and the translator is responsible for both post-editing and revising the content? In both these scenarios, there is one thing in common: the boundaries between translation, post-editing and revision is blurred. Konttinen et al. also point to another blurred line in the production process where “[a] parallel can thus be drawn between other-revision, modifying TM suggestions and post-editing MT output, particularly as these activities are increasingly integrated in the same software tools” (Konttinen et al. 2021, 193).

In all these cases whether post-editing replaces translation or not, the question of what is the knowledge and expertise expected from the post-editor need to be discussed in comparison to what was expected from the translator before TM and MT systems were involved in the process. Hence, the question: What about post-editing competences? Konttinen et al. state that “[m]any parallels to these revision competence models can be seen in various discussions of post-editing competences” (192). In terms of the second category, there is not much research into the issue of post-editing competences and only one in terms of ways to measure and compare the two competence models¹⁶. The two post-editing competence models are by Rico and Torrejón, and Nitzke and Hansen-Schirra. The earlier study grouped competences expected from a post-editor into three main categories, namely core competences, linguistic skills and instrumental competence. Under core competences, they place the two skills for “attitudinal or psycho-physiological competence” for coping with “subjectivity” in terms of text quality and “overcoming uncertainty” and also, “strategic

¹⁶ See Robert, Ureel and Schrijver 2022.

competence” to reach at ‘informed decisions’ (2012, 170). Under instrumental competence, they refer to “knowledge of MT systems and their capabilities”, terminology management skills, MT dictionary maintenance, corpus quality assessment skills, controlled language pre-editing skills and some programming skills (ibid.). Finally, under linguistic skills, the skill set expects the post-editor to have a translator’s expertise (unlike Pym) and “skills usually demanded of a translator”. These are “excellent knowledge of source and target language”, “familiarity with post-editing directions/guidelines”, “communicative and textual competence in at least two languages and cultures”, and, most importantly,” cultural and intercultural competence and subject area competence” (ibid.). On the other hand, Nitzke and Hansen-Schirra distinguish the skill set expected from three categories of people – post-editors, MT engineers and PE consultants – in workflow involving TM/MT in a comprehensive way. They categorize the competences each group of people is expected to have in a house-shaped diagram and in this “house of PE competence, the three pillars are framed by a roof, which represents soft skills for post-editors” (2021, 73). What is expected from a post-editor is shown in the oval part inside the ‘house’ with three skills: error spotting, error classification and error correction. Again, unlike Pym, what they suggest about expectations from a post-editor has a bilingual element to it:

Similar to translation and revision competence, a post-editor has to have proficient knowledge of the source and target language, as monolingual PE always bears the risk that content mistakes are not recognised (see §4.1.3). This basic competence is often referred to as bilingual competence in translation competence models (70).

But because post-editors have the chance of taking up the position of a consultant, “they also must have profound knowledge about risk assessment and service” in order to negotiate “which texts and projects can be post-edited and which not” with translation service providers and clients (76).

All in all, the revision and post-editing-related competences are summarized in Table 1 below:

Table 1: A List of Revision and Post-Editing Competences

PACTE (2005 from 2003)	5 sub-competencies and psycho-physiological components. Revision and editing is mentioned within the scopes of two sub-competences namely translation knowledge and strategic sub-competences.
Künzli (2006)	Focuses on syllabus design for revision, has a set of 3 competence categories. “Strategic competence involves revision-specific elements of defining the task at hand, applying relevant evaluation criteria, and deciding how to address a problem” (Künzli 2006, 11 cited Konttinen et al. in 192)
EMT Expert Group (2009)	proposes a hierarchical model of translator competence, which was later modified by the EMT Board (2017). In the latest model, five areas of competences defined are language and culture, translation, technology, personal and interpersonal, and service provision (EMT Board 2017)”.
Göpferich (2009)	A longitudinal study called TransComp is presented and set on the basis of a translation competence model of 5 sub-competences, namely communicative competence in at least two languages, domain competence, tools and research competence, translation routine activation competence and psychomotor competence (20-21).
Rico and Torrejón (2012)	Summarize the skill set of post-editors under three categories: core competences, linguistic skills, and instrumental competences.
Pym (2013)	proposes a set of ten skills for TM/MT age under three larger categories: “learn to learn”, “learn to trust and mistrust data” and “learn to revise translations as texts”. “[T]ranslator’s function can be expected to shift to linguistic postediting, without requirements for extensive area knowledge and possibly with a reduced emphasis on foreign-language expertise.”
Robert et al. (2017)	Test their hypotheses and TRC model with students by measuring revision quality. The new sub-competences they add to their model are the Knowledge about Revision, and Revision Routine Activation. The argument that “revising frame of mind as opposed to retranslating” is stressed.

Scocchera (2017)	Drawing from empirical data with students who took a course on revision. Presents her multicompetence model with 6 components: analytical-critical competence, operational competence, metalinguistic-descriptive competence, interpersonal competence, instrumental competence, psycho-physiological competence.
Nitzke and Hansen-Schirra (2021)	Distinguishes skill sets expected from 3 groups of people in a workflow involving TM/MT. Their book is one of the preliminary comprehension works that specifically refer to “post-editing competence”.

Putting the commonalities in all competence models to one side, the subcompetences that concern revision and post-editing can be grouped into three larger categories: (1) strategic subcompetences related to revision or the post-editing process, (2) interpersonal, psycho-physiological or attitudinal subcompetences, and (3) instrumental or tools subcompetences related to the use of translation technology” (Konttinen et al. 2021, 193). But in terms of the linguistic skill set under the strategic subcompetence category, there seems to be different/clashing points of view. Empirical studies discussed in the next part will shed more light on the topic.

After this brief overview into the current literature, the following part delves first into deeper detail for revision or reviser competences in the international standards and then the results presented by empirical studies. Therefore, the discussion highlights the boundaries and common ground between revision and post-editing from the practitioners’ and the trainers’ point of view. In the final part, a pivotal question is raised in light of the developments in translation-related technologies and the other stakeholder, i.e. the industry. Do the expectations of the industry align with the theoretical work on competences?

2.2 Editing and revising in ISO17100 and post-editing competence in ISO 18587

The competences the reviser and the translator should have is said to be the same by ISO 17100, with only a slight difference in terms of positioning in the workflow and experience. The reviser “who shall be a person other than the translator” (2015, 10) is expected to have

all the translator competences, the qualification expected from the translators and also, “translation and/or revision experience in the domain under consideration” (6). According to Mossop, the problem with experience is ‘vital’:

[R]evisers must possess the various kinds of knowledge and skill that come from experience as a translator, as well as experience in a specific workplace. In the absence of such experience, a reviser will not be able to recognize adequate translations or know when a change is needed (2020, 119).

The phrase ‘when a change is needed’ indicates that a competent reviser both needs to *correctly*¹⁷ evaluate the errors in a translation and correct them in the most fitting manner possible. In the case of post-editing, studies suggest that “it is essential that post-editors are skilled translators” (Nitzke and Hansen-Schirra 2021, 69-70). Therefore, the problem here lies in how to create a training method in which the novice translator has the chance to observe errors corrected by a professional working in the industry or someone with years of experience behind them, and also how to adequately and consciously correct parts of a translation even if the content repeats errors that are hard to annotate. The reason behind the problem of error annotation or revision is two-fold. Because “the use of TM matches is, in fact, a form of other-revision just like the revision of human translations” (Kontinen et al. 2021, 193), the post-editor might be asked to edit a range of errors from the easiest, almost statistical ones, to correct or revise other translators’ work, and a combination¹⁸ of the two. Thus, in order to compare and contrast the two sets of competences required by revision and post-editing, we need to first address where we position “post-editing” in the translation workflow, hence, where we have “the human translator”¹⁹ in the loop. Playing by the rule book, we need to understand the workflow explained by ISO 18587, which states

¹⁷ Evaluation of MT output and the correct or adequate evaluation is discussed in the next chapter.

¹⁸ ISO 18587 clarifies in a note that “the text to be post-edited can be the output of an MT system or the output of a combination of different sources (e.g., TM and MT)” (2017, 6).

¹⁹ ISO 18587 designates that the human (translator) in the loop in the stages of determining whether the source language content is suitable for MT by the TSP, pre-editing the source language content, informing the post-editor about “the estimated level of usefulness of MT output” and post-editing the content and post-editors’ providing feedback back to the TSP (2017, 5-6).

This document therefore restricts its provisions to that part of the process that begins upon the delivery of the MT output and the beginning of the human process that is known as post-editing (2017, v). [...] However, there is no MT system with an output which can be qualified as equal to the output of human translation and, therefore, the final quality of the translation output still depends on human translators and, for this purpose, their competence in post-editing (ibid).

It is clear from this statement that similar to revision in the ISO17100 workflow where the revisor has the final content produced by the translator, the post-editor has the (final) output produced by the MT system but also the accompanying source content. Therefore, the focus on just target language content shifts to a contrastive one that requires the post-editor to have critical bilingual competence as well. In this sense, we return to Konttinen et al.'s point where they stress that “[i]n syllabus design and pedagogical practice, a strategy and the domain where it is applied are intrinsically linked”. Acquisition of how to “arrive at informed decisions when choosing among alternatives” (Rico and Torrejón 2012: 170) depends on “knowledge related to professional translation practice” (Künzli 2006: 15) and on “knowledge about typical errors that need to be attended to when revising or post-editing” (Konttinen et al. 2021, 193). In other words, metalinguistic knowledge is what a training model seeks. But the knowledge of how, when and to what degree this knowledge is operationalized requires empirical evidence from different parties. The part about ‘what degree’ is largely based on the expected degree and type of errors and, hence, can yield assumptions which are not always accurate and in our case result in hyperrevision and unnecessary post-editing. Though, in the case of post-editing, “the distinguishing trait is the fact that the ‘agent’ producing the translation is an algorithmic machine” the post-editor can easily assume that “this origin of the translation is likely to be reflected in the features of the MT output, including typical errors” and therefore may not notice the major errors (Konttinen et al. 2021, 193-94). On this point, the training of post-editors needs to include some other skills compared to translators and revisers. On this point, ISO 18587 warns the trainers as it clearly puts forward,

The training language professionals in the knowledge and skills required for post-editing can provide language professionals with the required post-editing skills that are different from translation skills” (2017, 9)

The statements under the subheading “professionalism” of ISO 18587 refers to skills including general knowledge of MT technology and CAT tools, “how terminology management systems interact with MT systems” and whether it would be feasible to edit MT output in terms of time and effort, but also “a basic understanding of common errors that an MT system makes” and “the ability to follow the instructions received and the ability to focus on specific issues and make specific corrections as given” (8). This focus on ‘specific’ corrections implies that an almost mechanical approach is called for, depending on the purpose the content will be used for. Similar to the international standard, Konttinen et al. states

the main characteristic feature of post-editing in this group of subcompetences seems to be the ability to apply a set of principles or rules in choosing a post-editing level that suits the intended purpose. (2021, 194)

As part of professionalism, post-editors are expected to “provide structured feedback on frequently recurring errors in the MT output” and contribute to improving the MT system over time. Yet, these statements of professionalism might fall short if MT output requires the post-editor to adopt more than a simple mechanical approach as the aim of full post-editing is later stated as “to produce an output which is indistinguishable from human translation output” while it is still “recommended that the post-editors use as much of the MT output as possible” (8). So, it can be argued that it leaves an open door for the wide array of errors to be expected from an MT output. Moreover, the “list of what post-editors’ training can include” in the annex includes more advanced skills such as “advanced use of TM and MT technology”, “advanced terminology work”, “advanced text processing skills in order to be able to use macros and search and replace functions”, “practice in light and full post-editing”, and “use of quality tools to perform quality checks at the end of the project” (9). Because the standard lists the competences and skills post-editors should have at the end of training, the

micro steps taken during the process are not specified and need to be based on evidence and further tested. In other words, we need to delve into the details of what progress a training program aims to achieve.

Following in Donald Kiraly (2000)'s footsteps in distinguishing between translation and translator competence, it makes sense to trace the common points of strategic sub-competences of the above models and see how researchers interpreted and applied different methods to make prospective translators learn how to learn. After all, the acquisition of competences by the translator-to-be happens not via an input and direct output kind of a mechanism but rather a process. Therefore, it is crucial to shed light on the "role of the translator in the loop" from a social constructivist approach, alongside the evolving nature of the competence due to experience.

2.3 "Human Translator Competence": Describing translator's role in-the loop and the position of revision

The literature review and international standards related to revision and post-editing competences reveal what achievements are to be expected as a result of a training program. However, only a few of the studies provide information about the process upon which these competences are built. Because of this lack of empirical data about training, the common notion of building competences seems like a crystal-clear assembly line where each skill is added one after the other in a series of steps. But there is little evidence of how each skill and each sub-competence, for instance strategic sub-competence, is transformed into a training system in which problems are provided and solutions are asked for. With respect to this pedagogical shortcoming, Donald Kiraly's "post-modern", anti-positivist approach to translator education is pivotal for understanding how learning takes place in a vacuum in contrast to uni- or two-dimensional models that have "static box-like representations of an ideal(ised) relationship between dispositions, abilities and skills" (2013, 201). Kiraly's notion of "the emergence of translator competence", which refers to translation competence and translation competence acquisition, stressing the complexity behind these processes,

entails that “complex adaptive systems develop not through a simple or complicated (mechanical) accumulative process, but instead emerge fractally and in a self-perpetuating manner, yielding ever-greater degrees of irreducible complexity” (203). This shift from “translation competence” to “translator competence” not only shifts the emphasis to nonlinguistic skills, but also foregrounds how experience builds into sub-competences first, and then into what Kiraly calls “super-competence”; in his words:

[A]long with the volume of translation experience (including reflection and feedback), the complexity of the linkages throughout the system can be expected to have increased dramatically. In addition, the separate sub-competences would have merged into a highly integrated and largely intuitive super-competence” (2013, 208).

With the addition of this largely intuitive super-competence into the model, Kiraly’s four-dimensional new model illustrates each sub-competence “as a vortex to suggest constant evolution, ever-increasing complexity and the recursive nature of learning” (Risku 1998, 37 cited in Kiraly 2013, 212). The swirling, interactive vortices in his new model depict “an emergent process of progressively incorporating experience and learning into the emerging sub-systems” (ibid.). When he uses this multi-vortex model in a case study in which the teacher does not transmit knowledge, skills or competence in a didactic way, it opens a way for the embodiment of personal, authentic experience and enables the students to think ‘outside-the-box’ (214). In line with this post-modern direction in training, Senem Öner Bulut²⁰’s article titled “Integrating Machine Translation into Translator Training: Towards ‘Human Translator Competence’?” presents a case study designed to empower the ‘human translator’ during the training process with “awareness of their existing and potential roles as human translators” who will not be limited to post-edit and pre-edit the content provided in

²⁰ Öner Bulut provides a thorough list of all the competence models found in the literature which does not necessarily refer to post-editing competence: “Akalin and Gündoğdu 2010; Balkul 2015; Bengi-Öner and İnce 1995; Birkan Baydan 2011, 2013; Ersoy and Balkul 2012; Esen-Eruz 2008; İnce and Bengi-Öner 2009; Kelly 2005, 2008; Kiraly 1995, 2000, 2004, 2013; Kurultay 1997; Neubert 2000; Öner 2013; PACTE 2003, 2005; Pym 2003; Schäffner and Adab 2000; Tosun, Akın and Şimşek 2015; Yazıcı 2007; Yücel 2007” (Öner Bulut 2019, 3).

the workflow of a CAT tool but will also train MT systems, select and assess training data and collaborate with MT developers as expert human translators (2019, 1). Kiraly's stress on awareness is based on his social constructivist approach to translator training which emphasizes "professional self-concept" that is raised in the minds of the prospective translators in "a dynamic, interactive process based on learner empowerment" (2000, 17). In line with this argument, professional awareness is represented in the multi-vortex model as well as the concept of "translation meta-competence" that Öner Bulut analyzes with reference to Esra Birkan Baydan. Though Birkan Baydan does not clearly present a competence model with sub-categories, this meta-construct comprises all the elements of the models presented in the above section on a macro level. Her critique of on meta-competence, in particular, is paramount for the translator working in a TM/MT workflow:

None of these can be complete as if in a closed loop. In this context, perhaps the very concept of competence is misleading as it implies an acquired and gained skill or knowledge. . . . In line with such an understanding, translator competence is about how the translator positions his/her profession in his/her mind. (Birkan Baydan 2011, 168–169; from Öner Bulut 2019, 5-6)

Considering that there are more than one, two or even many more agencies working on the same content in this workflow, how shall a translator function in an 'open loop'? Quoting from Öner Bulut, "Is competence also becoming a shared competence as the case in agency?" (6) How shall a training model empower future translators and what specific steps should be followed²¹, assuming that training should be "pro-active" and "holistic," and in a context "in which translators have ownership, critical understanding and a good deal of control" (Kenny and Doherty 2014, 290)? Answering these questions, three studies are particularly noteworthy.

²¹ Calvo notes that not everybody agrees "whether translator and translation competence refer to the same notion" but "both concepts are frequently used interchangeably without further discussion" (2011, 7). Öner-Bulut and Kiraly examines this concept confusion in detail and also, state their approach very clearly. Yet, it must be expressed that Some researchers, including Fernández Prieto and Sempere Linares, think about translation competence as part of translator competence.

Starting with the earliest one, Dorothy Kenny and Stephen Doherty's article titled "Statistical Machine Translation in the Translation Curriculum", the researchers advocate for the use of cloud-based SMT instead of asking translator trainees and trainers to train their own SMT –because clearly training an MT system would be very intimidating considering that these systems are built for computer scientists, not for ordinary computer-literate people. They also propose a practice in a specifically designed syllabus in which students are asked to train an SMT system, select training data and make "judgements about the quality of training data based on their own linguistic and translation expertise and on criteria such as authorship and homogeneity of data, file mark-up (condition of internal tags), etc." (Doherty and Kenny 2014, 302). As a final step, they are asked to evaluate the MT output and provide "a typology of errors" from the output, "with a view to ascertaining whether there were any specific interventions they could make to eliminate some of these errors" (303). The authors highlight that as a result of automation, "the clearest change is in the concept of translation itself – not just in the names, but in the range of activities and the spread of the skill sets involved" (...). What additional skills might be necessary in the face of automation is a question both sides of training, i.e., the trainers and the trainee, must ask. In this respect, more recent studies by Joss Moorkens and Anthony Pym examine the impact of neural machine translation (NMT) on the curriculum and what "underrepresented" skills should be incorporated into training. Contrary to Doherty and Kenny, Pym and Torres-Simón (2021) stress that "automation-resistant skills" of human translators are what constitute the basis of translator training, referring to Justa Holz-Mänttari's definition of "interlingual communication" as "the expertise of the translator, their prime knowledge base and field of mastery", after discussing the types of skills NMT might "replace" (2) and how translators "might obtain appropriate social rewards" (14) for their fundamental interlingual communication skills. Moorkens' study, on the other hand, attempts to "demystify" NMT for students, "familiarise" them with it so as to "become aware that, despite the hype about machine learning, that [*sic*] NMT output has many weaknesses as well as strengths" (2018, 376) by integrating an evaluation exercise, which would contribute to their technological and

instrumental competence. Interestingly enough, Moorkens positions MT-related skills within the frame of translation competence (378).

Öner Bulut's study, mentioned above, follows a similar line of thought and proposes "the critical consideration of the emergence of 'human translator competence' as a novel construct of translator training (research)", stressing "the interaction between the human translator and the machine/technology might not be limited to the sub-competences" such as 'instrumental competence', 'professional and instrumental competence' or "technological competence", but rather as a competence that would "further lead to the emergence of 'human translation (meta-) competence'" (18). Therefore, Öner Bulut conducts a survey on translation trainees' opinions for the future of the profession, and the SWOT-analysis on this survey reveals "somewhat of a tension" (16). For instance, when asked what the impact of MT on the profession in future could be, a student comments that "we are going to try harder to be more competent than machines" which, in the author's interpretation, demonstrates that "the learning practice inspired a kind of self-reflection (16). Apart from self-reflection, "some degree of self-criticism" was observed in students' commentaries as a student compares her own "mistakes" and the "mistakes" of the machine and states that this comparison helped her realize why the instructor of the course does 'not understand' her translations at times". Thus, it is clear that use of MT in the classroom raised "student's self-awareness concerning her own translation 'mistakes'". Öner Bulut further argues that this learning practice not only helped students familiarize themselves with MT, but also induced the trainers to look for ways to train the students for the future and helped them "to develop a self-awareness, a professional self- concept as 'human' translators" (19). Parallel to this study, Öner Bulut refers to Maria del Mar Haro-Soler and Don Kiraly's (2019) research project on the use of MT in the classroom having a social constructivist approach and underscores that this specific approach is significant for "investigat[ing] constructs [self-confidence, self-esteem, self-efficacy (beliefs) and self-concept] related to the translator's psychological 'self'" with the collaboration of "teacher-researchers, student-researchers and student-subjects" and also for enhancing the "reposition[ing] of multi-faceted human translators"(cited in Öner Bulut 20). As an additional point, this stress on empowering translators "as agents" is also present in

Moorkens' study as cited from Kenny and Doherty 2014, which states that learning about NMT gives the trainees a chance to act in a workflow with NMT that would give the translator priority, instead of attributing them "a limited or reductive role" (Moorkens 2018, 4)".

Now, after this social constructivist approach to training, it is crucial to note that the expression "human-in-the-loop" in the sub-heading can also be interpreted as translators' exclusion from the workflow, probably due to automatic post-editing. Simply put, automatic post-editing can be defined as a method for improving MT systems using "knowledge extracted from datasets that include post-edited content" (do Carmo et al., 2021, 101). It is generally used for academic purposes in developing MT technology and has limited commercial use. It was a particularly hot topic in the Workshop on Machine Translation (WMT) between 2015 and 2019. Though in WMT reports it is not stated how data used in automatic post-editing competitions obtained from reference translations are collected (119), it is easy to guess that most post-editing segments in the corpora-in-question are taken from human translators' work. After WMT 2019, the subject of automatic post-editing was left to its own demise following the repeated disappointments in the results of these competitions in terms of the purpose it was first designed for. Because this study from start to finish focuses on the (human) practices in post-editing, there seems to be no need to describe this technology in depth. Hence, the rest of chapter will discuss mainly constructivist approaches to translator competence and training and the results obtained from exercises/experiments conducted with students or professional translators.

Another study that reports on hands-on experience working with MT is conducted by Joss Moorkens with undergraduate and PhD-level translation students (2018). Though it is not a controlled study by definition, and only an exercise for students to analyze MT output (7), the terms and its results are significant. Firstly, the main purpose of this research was to compare the two MT paradigms and for participants to express their preference for one paradigm. It was "intended to contribute towards the translator's technological competence (EMT Network 2017) and instrumental competence (Hurtado Albir 2007)" (4). The

participants were asked to choose a page from Wikipedia²² on a topic they found “familiar and interesting”, copy 20 randomly selected segments in a language supported by an MT system they choose and thence try to form two groups of ten segments of equal length. One group was translated by an SMT and the other by an NMT system. The output was clearly differentiated and copied to a Word document for evaluation. No CAT tool was used. Participants evaluated the texts based on three criteria: post-editing effort, e.g., total editing time, adequacy using a Likert-type scale and error typology including word order errors, mistranslations, omissions, additions and tip (6). They were also asked to measure average post-editing time per segment, the average segment adequacy score, the frequency of each error type “by dividing the total number of occurrences by the number of segments” (6). Following the evaluation and post-editing assignment, participants were asked to write a short report in which they were asked their preference for a paradigm, specifying the language pair, and to explain their reasons with examples. These reports were graded and awarded a course credit at the end. Moorkens provides the prior information for the exercise as seen below:

The most popular Machine Translation (MT) paradigm over the past 15 years or so has been Statistical Machine Translation (SMT). More recently, there have been claims that Neural Machine Translation (NMT), a statistical paradigm that carries out more operations simultaneously, can produce improved output (5)

The results of the exercise demonstrate that participants overwhelmingly favored NMT – 93% of each cohort – as is apparent from the evaluation criteria. Both cohorts reported that they spent less time post-editing the NMT output and the average rating for adequacy for NMT was higher than for SMT – 2.95 for SMT and 3.46 for NMT on the 4-point Likert scale. Participants also found fewer word order errors, fewer omission errors, fewer addition errors, and fewer mistranslations in NMT (7). A similar study by Castilho et al. found that “fluency

²² Moorkens expresses that given the time limitations of the task terminology searches were minimized by using and selecting texts from Wikipedia, even if they do not necessarily comply with Kenny’s suggestion for. using “appropriate, authentic texts” (Kenny 2007, 204 from Moorkens 2018, 6).

is improved, and word order errors are fewer using NMT when compared with SMT. NMT produces fewer morphological errors and fewer segments that require any editing” in the case of certain language pairs in the educational domain (Castiho et al. 2017a from Moorkens 2018, 7). But Moorkens states that “the authors found no clear improvement for omission or mistranslation when using NMT, nor did they find any improvement in post-editing throughput”. Another noteworthy point about Castilho et al.’s study is related to types of errors participants annotated, as she explains how they “found NMT errors more difficult to identify”, yet “‘word order errors and disfluencies requiring revision were detected faster’ in SMT output” (Castiho et al. 2017b, 11 from Moorkens 4). In Moorkens’ exercise, one student said that “types of errors [found in SMT output] were not difficult to correct” (8). Therefore, one additional advantage familiarity with NMT output gives to participants is that they have the opportunity to “improve the efficiency of NMT error identification” (4). In a detailed evaluation study comparing SMT and NMT by Popović (2017), the participants found NMT produced fewer overall errors, improved verb order and verb forms, with fewer verbal omissions, but was not equally successful “translating prepositions, ambiguous English words, and continuous English verbs” (3-4). One participant went as far as to suggest that NMT output was “of a quality sufficient to make monolingual post-editing feasible ‘by someone with no knowledge of German at all, removing the need for a German-speaking post-editor’” (8).

In terms of evaluation results for adequacy and post-editing effort, Castilho et al. (2017a) states the results were “inconsistent” and found “comparatively higher numbers of identified errors of omission, addition and mistranslation in NMT output”, so they argue “there is still much room for research and improvement” with NMT, especially in some language pairs and domains (110 from Moorkens 2018, 4). When evaluating the temporal effort, or time spent on post-editing, participants as well as other studies underscore that MT can be used as a translation aid, even though there is sometimes an issue of lack of enthusiasm from translator users (Plitt and Masselot 2010; Guerberof 2012)” (5). One final note about this study is that when students’ were asked about scenarios in which NMT use may be “appropriate”, they suggested “perishable texts, as a springboard for ideas during rush-jobs”,

yet for “highly inappropriate” cases they suggested “for high risk and literary texts, for translation jobs that involve regulatory compliance” (9). This argument leads the present study to ask whether the evaluated text, i.e. sustainability reports, is an appropriate case for no post-editing and revising jobs. The answer is given in Chapter 5 where we discuss the results of evaluation and the function source text plays.

Another significant study on human translator competence reveals some theoretical modelling on revision competence in contrast to translation competence and presents some very useful hypotheses enriched by their “revision-specific insights” (Robert et al. 2016, 2). This study also presents empirical research done with participants from translation trainees. Because what post-editing competence requires can be better explained in comparison with revision competence, and also, “the ability to revise translations as texts” is a new skill for a new translation competence in the MT age, theoretical work into translation revision competence (TRC) models needs to be tested and validated. So, Robert et. al. start by pointing out that their study is based on validated models from “an empirical-experimental perspective” (1). Additionally, their TRC model is part of “a complex construct consisting of different subcompetences, that is, as a multicomponential competence” and does not consider translation competence as ““summation of linguistic competences” (1). Stressing the fact that there are not many studies on the specific skills revisers should possess apart from those of translation competence, the subject matter of their model focuses on what Mossop called “‘a mental switch’ from a ‘retranslating’ to a tolerant ‘revising’ frame of mind” (3). They also highlight that the literature review into the concepts of revision and competence discloses “a halo of fuzziness surrounding the concept of revision” (3). But they argue that “the term revision should apply only to the revision of a translation by a reviser who is someone other than the translator, carried out before it is delivered to the client” (4) and hence, they base their model on Mossop’s definition of revision, which is:

Revising is that function of professional translators in which they identify features of the draft translation that fall short of what is acceptable and make appropriate corrections and improvements. (2007, 109)” (4)

In terms of competence, their model is based on Hurtado Albir's definition of competence (2007, 166), which, simply put, refers to understanding of the requirements of a certain job. Therefore, their TRC model is built on "two multicomponential TC models that have already been empirically investigated" with PACTE and TransComp's translation competence models (4). Robert et al. explains their choice of PACTE's definition for competences as the group refers to cognitive psychology where "a distinction is made between explicative knowledge (know why), declarative knowledge (know what), operative knowledge (know how) and conditional knowledge (know when and why to use knowledge" (PACTE 2011, 339 cited in Robert et al. 2016, 7). Again, referring to findings of PACTE, the foundation of TRC lies upon the validated data suggesting "dynamic knowledge of translation is a characteristic of TC acquisition, with a progression from static to *dynamic declarative knowledge over the years*" (PACTE 2014, 109, cited in Robert et al., 7, stress mine). In line with this, Robert et al.'s point about Göpferich's TransComp TC model is that "the more complex strategic subcompetence does not develop until less complex subcompetences have reached certain threshold values" (Göpferich 2013, 74 from Robert 2016, 9). EMT's proposal for competences, on the other hand, has two dimensions: an interpersonal dimension and a production-orientated dimension, and under production-orientated dimension, "EMT expert group briefly mentions the necessity to know how to proofread and revise a translation (mastering specific techniques and strategies)" (9-10).

After laying the foundations of TRC on these earlier macro-level translation competences, researchers summarize the "crucial elements of a reviser's TRC" as "(1) translation experience; (2) interpersonal aptitude; and (3) the ability to justify changes to the text" (10). The primary question here with the first element is whether all revisers are always experienced people. The answer supported by surveys in earlier research says no. Robert et al. refers to Sager 1994; Rochard 2002; Gile 2005 with revisers being experienced translators. The EMT report written by Schjoldager, Rasmussen, and Thomsen in 2008 also supports this point where "the work (editing/revision and précis-writing) is primarily done by trained translators, but also by people with other educational backgrounds". Yet, Sedon-Strutt (1990) argues that experience as a translator "is not an absolute requirement for (high-quality)

revision” and suggests that there are advantages to “having junior translators revise the work of senior colleagues”. Mossop’s study (2007) also confirms this latter point (cited in Robert et al. 2016, 10). Finally, Hansen conducted a longitudinal study on revision competence and asks:

[w]ith respect to the question of ‘experience’, it is interesting that the two experienced translators were already good at both tasks in 1997. As they were competent 10 years ago, the question could be asked: what has ‘experience’ contributed to their revision competence? (2009, 269)

However, it must be noted that these studies are not empirically validated. Thus, the ISO 17100 requirement for revisers to have “translating experience in the domain under consideration” besides translation competence is not always fulfilled and the door is open for empirical research on this subject matter. When it comes to interpersonal aptitude, studies list a number of qualities revisers should possess, such as “judgement, an open mind, sociability, respect for others, honesty, modesty, patience, a sense of responsibility and organisational skills” (Horguelin and Brunette (1998); for Mossop, interpersonal relationships; for Hansen, “fairness, tolerance” (11). The final dimension of the TRC model, i.e., the ability to justify textual changes, is a skill that almost all those named in the review mentioned. For instance, in his proposal for a revision course, Mossop suggests three directions that would be helpful as goals of a revision course:

(1) professional preparation for work as a translator, i.e. preparation to be revised, to self-revise²³ and to inter-revise; (2) preparation for target-language editing work; and (3) the theory of rewriting (Mossop 1992 cited in Robert et al. 2016,11).

²³ Mossop suggests that reliance on self-revision, put differently, “over-confidence in one’s ability to avoid error” observed especially in cases of highly experienced people, has its disadvantages as well. The most apparent one is that fewer errors will be detected in any work since “translators have a certain blindness to their own translations” (Mossop 2020, 121).

Apart from the final goal on theoretical training, these goals target both achieving the mental switch and learning how to justify changes. This brief summary of revision goals points to aspects of post-editing as well. In this sense, both activities seem interchangeable. This issue will be discussed in the overview so at this point we leave it here to return to three hypotheses underlying the TRC model, with regards to three aspects found in the literature: (1) the definition of the revision process, (2) the definition of revision competence and (3) the subcompetences of the revision competence. For the first hypothesis, the study assumes that the translation process consists of three main operations, namely “(a) reading the target text (with or without the source text, in function of the revision brief) for comprehension and evaluation, (b) detecting problems by applying a detection strategy²⁴, (c) applying either an immediate solution or an immediate or postponed problem-solving strategy”, which includes “rereading, reflection-reformulation, search operation” (12). The second one is related to the revision operation per se, stating that “the translation revision competence is the underlying system of knowledge needed to revise someone else’s translation”. This specific competence requires one to carry out the revision by “taking into account the purpose of the translation and the revision task”. The final hypothesis assumes that TRC consists of strategic subcompetence, tools and research subcompetence, interpersonal subcompetence and revision routine activation subcompetence (12-13).

The pilot study was planned to be conducted with translation trainees who took the revision course (cohort 1) which was designed on the TRC model and those who did not (cohort 2). The study does not give the details of the result because the pilot study is in its proposal stage. Yet, the researchers expect certain outcomes: that those who took the course will have acted differently in comparison to trainees who did not take the course. For instance, they expect all participants to be “able to work with (some of) the necessary tools since they are essentially tools that are also used for translation” but because the frequency with which the translators use tools for revision is much lower, they expect cohort 2 to be less efficient and to have much less practice in using these tools. As for revision

²⁴ This detection strategy generally involves “either an operation of anticipation of potential problems or an operation of comparison between an item in the text and knowledge stored in memory”. It must be noted that knowledge may be certain, uncertain or lacking in many respects (Robert et al. 2016, 12).

subcompetence, the researchers “expect translators to have no or very little declarative knowledge of revision”. Furthermore, in the case of strategic subcompetence, they “expect revisers to apply the most appropriate revision procedure and make only necessary changes, which is one of the major revision principles”. One additional point is related to fairness and tolerance. The researchers expect cohort 1 to “make fewer categorical statements about suggested changes when exposed to a revision task without clear revision instructions” (15).

One more study analyzing the attitudes of the people on each side of the coin was conducted by Di Wu, Lawrence Jun Zhang and Lan Wei on the trainers’ beliefs, training practices and competences. The researchers start by addressing a gap in research regarding the 21st century skills and competences of translator training, including instrumental competence, professionalism and ethics and revision competence, which they believe are “significant components of the training needed to better prepare translation graduates to meet the needs of the current language services industry” (2019, 233-34). Because they work in the Chinese context, the insights they gather from trainers is in agreement with international institutions in that trainers in China agree that “translator training should not be limited to developing language-related competences of trainees (e.g., PACTE 2003; Kiraly 2000, 2006, 2015)” (234). Yet, they suggest that there exists “a gap between the pedagogical recommendations derived from translator training research and their uptake in training practices”. They propose that this gap might be because ““deep-seated epistemological assumptions and beliefs’ (Kiraly 2006, 78) in the trainers has led to a lack of alignment between academic research and the actual teaching practices of translator trainers” (234). Reviewing the competence lists and models, they surmise “the concept of translation competence is either expanding or not yet stable” (235). They refer to a lack of empirical research on this subject matter and assert that most research is “based largely on propositions or anecdotal experience of translator trainers or practitioners” and hence, has been considered “simple speculations” (Lesznyák 2007, 172 cited in Wu et al. 2019, 235). But on a positive note, at least these speculations “serve as theoretical bases for empirical studies” (235). The progress of how competences are illustrated indicates “a change in epistemology and both view translation as a cross-lingual as well as a vocational activity”, pointing towards “an

expansion of translation competence” into what studies term “translator competence” or the new inter-personal dimension of a multi-componential model and professionalism (236).

In order to observe this change in epistemology, Wu et al. used a mixed-method approach with data obtained from translation trainers using questionnaires, classroom observations and interviews (238). Data obtained from 79 participants and analyzed statistically demonstrates that “the translator trainers believed that all the sub-competences were of moderate importance for the development of translator competence” (242) and “bilingual and translation knowledge sub-competences were the major foci of the participants’ training practices” (245). However, when the questionnaire results were compared with interviews, the analysis revealed that “there were some discrepancies between the beliefs and training practices”. After asking the participants to explain these discrepancies, the interviews, surprisingly, disclosed that the mismatches were attributed to a variety of internal and external factors, including trainers’ self- efficacy and lack of motivation. For instance, some participants admitted they had “low self-efficacy in technology” (245). The external factors, on the other hand, covered “students’ abilities, curricula, examinations, colleagues and facilities”. The major external factor noted by all the participants, was students’ abilities, by which they meant students did not reach a high level of language proficiency and therefore, trainers found it “unnecessary” to focus on the other sub-competences (246). Wu et al. concludes by suggesting three “implications for the (self-)training of translator trainers”. First, trainers can benefit from “competence-specific and trainer-oriented” training; secondly, “initial translator trainer education” can be very beneficial for trainers, and lastly, translator trainers can also benefit from “self-training”, which the authors explain thus:

By training themselves, translator trainers may become able to deal with specific issues in their own teaching contexts which are not resolved in translator trainer training. Translator trainers need to be life-long learners by being actively involved in many activities, including undertaking professional translation tasks, reflecting on teaching, communicating with colleagues, conducting research and attending academic conferences, as mentioned by the participants in this study (247-48).

The interaction between translation trainers and trainees, as well as life-long learning, is the subject of one final work with a planned structure and detailed results presented in Giovanna Scocchera²⁵'s study with students on revision phases. The author notes that this study “was designed and implemented as part of a Translation for Publishing Purposes module” and its purpose was to “investigate the students’ general attitude towards revision and their background revision skills, and how the short-term course affected their revision competence acquisition” (2). Referring to Hansen’s theoretical work on the attitudes and skills necessary for revision (Hansen 2009), Scocchera stresses the fact that apart from attentiveness to linguistic and stylistic issues a reviser must be fair and should have the “the ability to explain and argue for changes” (Hansen 2009, 323 cited in Scocchera 2019, 3). It’s no surprise that she underscores the argumentative and declarative competence and positions them at the core of any training experience. As a result, trainees should have a knowledge of “relevant expressions (terms) they can use professionally” in order to evaluate, clarify and “justify” revisions on texts which are in most cases translated by a colleague or retrieved from TMs (Hansen 2009, 313 cited in Scocchera 2019, 3). Thus, she advocates using evaluation metrics and sample errors as useful aids for eliciting argumentative and explanatory skills in the proposed revision course, which should have “a learner-centred approach, where the teacher acts as learning mediator and facilitator in group discussions and exchanges” (4).

The training model she proposed asks students to complete a questionnaire at the end of the course so that the quantitative and qualitative research methods could be blended (5). In short, students were given four separate revision tasks at successive stages, including an editing task, a comparative revision, and two subsequent comparative revisions accompanied by comments written before and after the course content. The same translated text was used in all the tasks, and tools for revision were provided to participants who were, in fact, students who took the course. The participants were divided into two groups. The first cohort was

²⁵ Studies on revision competence tend to explore the concept as a theoretical product and not focus on anything applicable, from training contents to material, methods and tools. In addition, its place in curriculum and the course design are also among the subjects lagging behind.

asked to track changes made to the text in all the four stages with revision tool on, whereas the second cohort was asked not to track the changes with revision tools off. The crucial point in all these stages is that students were given the freedom to decide whether to translate from scratch or use the revision work they were assigned for the new task. Collaborative work was not accepted and therefore, students were instructed to “address an imaginary translator when suggesting changes or writing comments”, acting as individual freelance translators (5). The authors state that the study targeted some general objectives and some other stage-specific ones, such as exploring the students’ attitude towards revision and their background skills and knowledge in revision; looking into the level of expertise acquired in each revision sub-competence; testing whether course content and methods were valid and efficient; observing whether training has a positive impact on revision competence acquisition; collecting qualitative and quantitative data on the relationship between different revision procedures and tools, and revision products (6).

In the first stage, before any course content is provided, participants are asked to perform monolingual revision or editing. To create a natural experimental set-up, they were left completely on their own in terms of the level of revision work. The purpose of this stage was to produce data for studying the relationship between tracked and non- tracked revision outcomes in terms of type and number of changes introduced and observing whether a specific pattern can be identified as a revision trend (6-7). Different actions on the text were classified and analyzed using the taxonomy of revision changes, i.e., Faigley and Witte’s (1981), adapted to the study’s needs. The categories of errors were subdivided into these headings: “introduction of formal errors” for spelling errors, “addition” for connectives and repetitions, “deletion” for tiny grammatical entities such as subject pronouns, possessive adjectives, redundancies, “substitution” for word and idiom choice, register and verb tenses, and finally, “permutation” for changes in adjectives and/or adverbs in order to avoid calques (7). The results for this stage confirm the initial hypothesis, pointing to “a greater feeling of freedom and power”; non-track changes students (cohort 2) felt as if they had performed a higher number of revisions, while the cohort 1, the ‘track changes on’ group, might have felt the hypothetical translator to be a “restraint” (8). The track changes off groups applied

substitution extensively – twice as frequently as the ‘track changes off’ group –, their average number of additions and permutations were higher, but their average number of introduction of formal errors and deletions were lower. On the contrary, the ‘track changes on’ group had a higher number of spelling errors introduced to the text compared to the ‘track changes off’ group. The authors suggest that this might be related to the predominance of the color red in the file and the participants’ inability to properly differentiate the text from comments. For trainers, these results are interpreted as an objective for training in which students should be made familiar with the practical and ethical issue of editing, including “editing risks” such as “error introduction, misinterpretation and ultimately rewriting and retranslation”; hence, they have a clearer understanding of their freedom as revisers to change the text (8). Also, in terms of typing errors, students must get to know that they can change the revision settings so the color they use for changes stands out more easily from spelling error underlining.

In the second stage, both cohorts were given the source text and asked to do a comparative revision. Again, they were free to use editing work previously done or to start afresh. The hypotheses they tested at this stage were related to reversed revision actions and new ones. Because both cohorts had access to the source text, the non-track changes group would be expected to have a lower number of substitutions and changes and they would also have a higher number of reversed changes (9). In cohort 1’s work, “a dramatic drop” was observed in the average number of permutations. In cohort 2’s work, there is “a marked reduction” in all categories but substitutions still score higher. But the hypothesis for reversed changes in the track changes off group was not confirmed and these new types of changes were in fact “reduced to a minimum in both groups” (11). What is striking at this stage is that participants proposed suggestions and alternatives as opposed to “straight-forward substitutions”, suggesting that comparative revision gives the participants a larger view of context and wider range of translation solutions. The low number of new permutations, on the other hand, indicate that both cohorts only aimed for structural calques. In this sense, it can be argued that even with just monolingual revision errors in translation can be detected and repaired. Moreover, having a macro-overview of the translation even in fact proves to be an asset as the source-target comparison requires a micro-view of the text and causes tiny errors

to go unnoticed (11). One final piece of data concerning the higher number of reversed changes from participants who had a number of substitutions in stage 1 supports “one of the golden principles of revision work whereby ‘less is more’”, which is advocated by Mossop (11).

In the third stage of the study, participants were asked to review the revision work done up to this stage by adding comments. The specific objective of this stage was to see whether there were any patterns to be seen between the comments and the revision actions before any training regarding evaluation manner was given. The hypothesis at this stage was that because the reviser must spend more time thinking critically about the comments, new revision changes could be made and since the participants were not given any training, they would not be able to communicate all their thoughts and hence, the number of comments would be limited (12). The author categorizes all the comments made under five headings: “comments showing some level of metalinguistic awareness”, “partially justifying comments”, “subjective comments”, “judgmental comments” and lastly, “diplomatic comments” (12). Both cohorts commented on the same issues – “calques, word order, register, lexical accuracy, idiomatic language, readability” – and no significant difference between groups was noted here. Students who were more active in stage 1 continued to be so in the rest of the study. With respect to this, the authors argue that some revision patterns emerge from data obtained through comments, in other words “individual ‘revision styles’ and approaches do exist” (13). Also, because the issues raised were quite similar to stage 1 and 2 participants might have felt a tendency to perform more revision actions. These new cases of revision activity might be a cause of commenting activity as well. But the low number of comments at this stage can also be interpreted as participants’ “limited competence in formulating accurate and effective comments”, and the changes made may not be as eye-catching as the ones made in the first two stages (13).

In the final stage of the study, towards the end of the ten-week course students were introduced to several taxonomies including Mauriello (1992), Mossop (2001), Berman (2003) and afterwards, they were asked to comment on the revision work they commented on in stage 3. The purpose of this final stage was to compare the number of comments, as

well as their formulations of comments using the metalanguage they were recently trained in. The results confirm that there is “a general downward trend, albeit not particularly marked, is found in both subjective and judgmental formulations” (14). The author interprets this as a confirmation of the initial hypothesis which stated “revision-specific training produces enhanced commenting, metalinguistic and interpersonal competence” (14). Yet, the results signal that interpersonal competence could be more easily improved than metalinguistic competence because mastering the categories required a longer period of time. There is an observable change in terms of register in the comments when compared to comments written in the previous stage and quite surprisingly, in this stage of the study positive feedback was given to the hypothetical translator and the students started to give some kind of revision summary in their comments (14). Students also mentioned objectivity as a positive asset they gained from this task as they limited their actions to just those of reviser rather than acting like a second translator, which can also be linked to the track changes on and off setting (15). Some students suggested that having the tool on made it much easier to reverse or reconsider the changes made in the previous task and further on, when they were asked to write comments participants found it easier to justify their changes thanks to declarative and explanatory activity (16). Hence, among the other assets gained from this task they listed “increased critical awareness, enhanced vocabulary to explain the rationale behind their revision choices, and a new attitude towards the reviser-translator relationship” (16). When asked which part of the study they found the most difficult, most students chose comparative revision and revision with comments. The author states that comparative revision requires rethinking and reconsidering changes previously made and the comments call for the participants to have the right vocabulary for evaluation, and also, “the fear of ‘subjectivity’” is apparently a big problem. After the training, once they overcame these problems the majority of participants found the final revision task easiest (16).

When asked for feedback, students suggested more practice on the analytical-critical stage of revision work and application of metalanguage, more revision practice, more individual feedback on revision assignments and more practice on interaction (17). 19 out of 22 participants also said their attitude towards revision clearly changed. Overall, demonstrating

the difference between editing and comparative revision with a hands-on approach verified its effectiveness in cultivating analytical and critical attitude in different procedures of editing activity (18).

One final study, which focuses primarily on problem solving, regarding the concatenation of a determinant to the present study – the determinant here is experience – was conducted by Izumi Inoue and Christopher N. Candlin (2015). Their task-based experiential learning program was designed to make students familiarize themselves with industry requirements “beyond the mere translation of a text” and their main argument advocates “making students aware of novice-professional differences in the identification and solving of key challenges” by making students get involved in “authentic and contextualized tasks reflecting a wide range of real-life challenges” (71-72). Thus, the final question of this part is related to communication in a post-editing task. Revision clearly requires communication between human agents, whereas post-editing is conceptualized as a process between MT output and a post-editor and probably some type of interactive tool. Consequently, the main characteristic quality related to post-editing in terms of interpersonal, psycho-physiological or attitudinal subcompetences is the ability to apply “a set of principles or rules in choosing a post-editing level that suits the intended purpose” (Kalle Konttinen 2021, 194). Yet, it is very obvious, as it is also noted by Jakobsen, that “(post-)editing is considerably different from revising a human translator’s draft translation” and requires different skills from ‘ordinary revision’ (2019, 74). In that case, we need to understand the industry’s requirements from post-editors, which is the topic that will be discussed in the next part.

2.4 Do the quality expectations of the industry correlate with students’ expectations?

The subcompetences and competences regarding revision and post-editing have been discussed so far from either an academic perspective with a stress on training or from the standpoint of students/participants taking part in a study. Yet, research into what is expected from post-editors from the other main stakeholder, i.e., the industry, is lacking. Therefore, the critical review article by Valentina Ragni and Lucas Nunes Vieira on what to expect from

NMT after the significant shift in terms of MT architecture is also noteworthy as the authors elucidate the key points of literature on post-editing studies. The authors searched for “neural machine translation” or “NMT”, “human”, “post-editing”, “PEMT”, “assessment”, “evaluation”, “perceptions” or “attitudes” in the query on Google Scholar, Web of Science/Web of Knowledge (WoS/WoK), Scopus and ArXiv.org for any kind of publication published between 2015 and 2019. Four key trends emerge from this corpus of literature: the majority of studies focusing on quality and human evaluation, post-editing and productivity-centered studies based on processing time or throughput, perceptions of students and professionals, and investigations into stakeholders including the end-users such as practicing doctors, which are very limited in number.

As to quality and human evaluation, most studies assessed quality by empirical means, for instance by performing some form of linguistic or error analysis but addressing quality through non-empirical means was also a major topic of discussion. Manual quality evaluations with comparative side-by-side ranking and rating using various scales were very common. Yet, the results obtained by human evaluation were “less clear-cut” and sometimes contradictory, while the strongest results in favor of NMT architecture in comparative evaluations with SMT and PBMT²⁶ were based on automatic evaluations. This fact underscores the importance of human evaluation methods to get a much more detailed understanding of the merits and pitfalls of NMT (6). In this respect, a number of studies indicated that both students and professionals considered NMT errors more difficult to identify (Moorkens 2018 and Castilho 2017b from Ragni and Vieira. 2021, 7). These studies yielded results which shed light on what training should include for translators to use it effectively so that no errors in the target text go unidentified, i.e., “unedited”. These results include successful cases of sentence reordering, inflectional morphology and subject-verb agreement, and also weak cases of declining quality when sentences get longer and other

²⁶ There is one exception to this generalization. In Castilho et al. 2017b (114), more error-free segments were found in SMT (25%) than in NMT (2%) but the NMT output was found to contain fewer overall errors compared to SMT in some other studies (Wu et al., 2016; Moorkens, 2018; Castilho et al., 2017a). Castilho also reports that Increased adequacy (or accuracy) and fluency scores for NMT also increased. However, accuracy scores do not necessarily demonstrate a similar pattern.

semantic issues concerning rare words as well as lexical errors including omissions, sometimes up to four words, and mistranslations (7). Therefore, the results clearly demonstrate that semantic issues above the sentence level, for instance those regarding negation and prepositional phrases require the context to be examined over multiple sentences at a wider discourse level. Document-level approaches strive to fill in this gap in evaluation where terminology errors and omissions are turned a blind eye to due to evaluation limits. Because the studies that involve some type of human evaluation suggest a considerable increase in terms of fluency and the output reads fluently, missing content goes unnoticed in monolingual evaluations (7). Put simply, “what one system architecture struggles with, the other can often address well” and in order to reduce the required post-editing effort, what is needed is a combination of MT approaches (7).

When it comes to post-editing effort, the difference between NMT, SMT and RBMT was not significant enough to suggest that one system outdid the others in all respects. In terms of technical effort, a study found that NMT involved an increase of keystrokes per word in certain text blocks, indicating higher technical effort but Castilho et al. (2017b) state that even though the technical effort was reduced in NMT the results show mediocre improvement. On the other hand, in terms of cognitive effort both the number and length of pauses were highest for NMT compared to SMT and RBMT, suggestive of a higher cognitive load. Overall, the results point to “just a partial consensus on reductions” in post-editing effort with NMT output and do not yield a specifically higher quality output requiring lower effort²⁷ associated with it (8). Productivity, in similar fashion, is measured in terms of temporal effort, e.g. words edited per hour or time spent on post-editing. However, when asked about their productivity with NMT and other MT output, the authors had “a broader conceptualisation based on more than just speed or throughput” besides post-editing time, edit distance and number of characters edited (8). There is even one study where NMT “does

²⁷ Research on NMT post-editing demonstrates that similar to research on MT architectures before NMT, unnecessary, preferential changes still have a place in post-editing of NMT output even in studies where participants are explicitly asked to correct only meaning and grammar errors (8).

not seem to boost productivity as much as might be expected” (Sánchez-Gijón et al. 2019, 55 from Ragni and Vieira 2021, 9).

Another parameter which has a direct effect on productivity is the level of experience the participants have. 40 percent of all the studies in the above-mentioned corpus had professional translators as participants who were asked to translate from scratch, post-edit MT output, revise post-edited versions, or assess quality. 22 percent of the studies had professional linguists who did not necessarily have specific translation experience but had relevant linguistic knowledge – language teachers and academics. Interestingly enough, in 23 percent of studies laymen (bilingual or monolingual) were asked to take part (10). The objectives of these studies were to investigate expectations towards NMT, quality expectations and errors expected to be more specific. Some others tried to assess reading comprehension for texts translated using SMT and NMT with university students and staff, to test MT efficiency (11). Thus, all the studies assumed participants had either acted as “quality evaluator” or “technology user”. The objective of the studies changed in accordance with the assumed roles of the professionals as the “ultimate beneficiaries” of an MT system though the assessed tools and the evaluation methods changed (11). Still Ragni and Vieira note that studies examining “end-users of NMT (both professional translators and lay users) are lacking, and merit future work” (11-12). In line with this, the main distinction between experts and non-experts, with professional translators as experts and anyone who has some degree of foreign language knowledge from students to bilinguals, who might not have any translation experience at all as non-experts, is the independent variable of most studies. Predictably, Ragni and Vieira stress the fact that “[t]here is now clear evidence that, when these two extremes of experience are considered, evaluation results for NMT can dramatically change” – referring to Toral et al. 2018 – “especially when looking not just at individual sentences but at entire documents” (Läubli et al. 2018) (12). Particularly when MT output is fluent and it is harder to spot the mistakes, laymen can be more tolerant of them or may not notice the mistakes the professional can easily spot. On that note, the argument

continues with the shelf-life²⁸ of content because if those unnoticed errors do not matter, from the industry's point of view, does the content still need to be corrected? This question is the topic of the next chapter; therefore, it will be left unanswered here, but the authors of the article stress that they noticed "a gap in the reviewed corpus with regards to relevant questions concerning whether end-users are becoming more tolerant of errors because of (N)MT developments, or how, relative to other architectures, NMT might affect communication between its end-users" (12). First claimed to be a "disruptive technology" by many in the engineering sphere of the industry (Shoshan 2018), NMT and generative AI and other similar large language models have not replaced human translation yet. But they have certainly changed the nature of the editing task and the challenges initially speculated upon remain the same: we still need to train people and "find suitable points at which translators can intervene in or control (...) the translation process, and to educate new generations of translators/post-editors who are capable of working with NMT in a sustainable way" (Kenny 2018, 67 cited in Ragni and Vieira. 2021,13). The authors note that omissions and "fluent-sounding but semantically inappropriate" translation choices (Koehn and Knowles 2017, 6) are on top of the list of the topics related to NMT output evaluation that the trainees should have some experience in to acquire competence. As a consequence, the authors highlight that monolingual target evaluations and sentence-level evaluation, as opposed to document-level, cause flaws to remain undetected and, hence, using other methods will potentially expose them as the trainees gain familiarity with NMT output (13). The authors also express caution about automatic evaluations. For instance, hTER, which measures edit distance compares the final post-edited product to a reference translation to calculate productivity and does not measure every edit operation and, as a result, ends up penalising equally valid renditions that do not match the reference closely (Marg 2016, 25 from Ragni and Vieira 2021, 15).

One last point the authors make about their corpus is related to a lack of studies examining "the relationship between these stakeholders' views and their capacity to exploit the potential of this new technology whilst also being aware of its limits (Esperança-Rodier

²⁸ "As Ragni and Vieira explain, with the increasing requirement for 'fully customisable, dynamic levels of quality' (Way, 2013, p. 2) in the industry, NMT errors may not have any repercussion in cases where the shelf-life of the content is short-lived and the stakes are relatively low" (12).

et al. 2017, p. 12)” and “scarcity” of in-depth interviews and questionnaires involving professionals and students (13). Therefore, the reactions and impressions of both experts and non-experts provide valuable data, not just in terms of workflow and the project management side of the translation process, but also in terms of post-editing where specialized knowledge “may be even more urgently needed with NMT” (13). For both groups of people to “act as authorities” – the notary function’ in Pym’s words – of post-editing, training needs to address both the groups including the professionals who are not used to working with NMT as studies measuring “the actual usefulness of NMT for professional translators are still very scarce” (Forcada 2017, 305 cited in Ragni and Vieira 2021, 14). Given this scarcity in training how both groups trust or mistrust the MT output is addressed as part of the questionnaire assigned before the post-editing part of the present study. Assessing the post-editing behavior, particularly of novices given their perceptions, will give us clues not only towards what training should include but also on their revision competences and communication thereof to others. The next chapter deals with the post-editing behavior of experts and non-experts in more detail.

2.5 Overview: Do social constructivism and stone-hard truths of the labor market align? or human at the core vs. human in the loop

This chapter summarized the approaches to competence in translation studies literature and international standards. It followed the evolution of the concept from its initial formulations to current discussion on it, integrating post-editing competence into the social constructivist take on translator competence. The multi-vortex model proposed by Donald Kiraly has probably never enjoyed this much discussion in terms of incorporating skills related to the post-editing of MT as it is one of the most crucial topics on the agenda of translator trainers than ever. How to encourage translation trainers to acquire linguistic sub-competences including knowledge of error taxonomies and how to interpret them as well as interpersonal sub-competences for voicing and explaining their decisions is a complex matter depending on who the training addresses and the training environment. As time passes, translation

environments become increasingly merged and “the distinction between TM and MT is becoming more blurred” (Moorkens, 2017). Thus, empirical investigations into how translators are actually working with NMT and TMs are much more relevant with only a few studies²⁹ working their way into it.

To conclude, the labor market of today demands a novice translator to quickly gain experience in a certain domain, making it hard to profit, let alone even make a living, from unspecialized domains. Considering that specializing means efficiently operationalizing context and text knowledge into a tangible content in the minimum amount of time possible, we need to look at how the content already available, i.e. the TMs, fuzzy matches and MT, has a “priming effect” on a group of translator and trainees. The literature on MT evaluation by numerous groups of people with varying degrees of expertise is wide. An OECD report titled “The demand for language skills in the European labour market: Evidence from online job vacancies” states

Language proficiency is a form of human capital. The returns to language proficiency in labour markets tend to increase with the number of speakers, vary depending on contextual factors, other forms of human capital individuals possess and the unique combination between individual and contextual circumstances (Chiswick and Miller, 2003[8]; Chiswick and Miller, 1995[9]) (Marconi et al. 2023, 8).

Individual translation behavior and contextual circumstances are also part of post-editing behavior. But the relationship between the priming effect and post-editing have not been explored in detail. In the next chapter, the present study discusses the output of different levels of editing in relation to the priming effect and expertise.

²⁹ In the corpus of studies compiled by Ragni and Vieira, only the following ones provide the results of such empirical work: Sánchez-Gijón et al. (2019), Daems and Macken (2019), Läubli et al. (2019) and Zaretskaya (2019).

CHAPTER 3 PREFERENTIAL CHANGES AND EDITING

THRESHOLD: EXPERTISE IN DOMAIN AND THE NOVICE

TRANSLATOR

One of the earliest definitions of “post-editing” was given by Harold Somers as “tidying up the raw output, correcting mistakes, revisiting entire, or, in the worst case, retranslating entire sections” of a machine translated text (2003, 138). Since the early days of CAT tools and statistical MT paradigm, much has changed both in industrial settings and text-centered research in TS. This chapter is devoted to a comprehensive literature review on both types of post-editing, full and light, with experiment results. What is crucial for interpreting the results is the concept of evaluation and, in this sense, a brief discussion of evaluation from the notion of “optimal equivalence” to the present-day frameworks and guidelines for examining whether a translation is ready for the market follows. A *sine qua non* discussion on preferential changes and inter-annotator agreement will thence ensue. Research results focusing on different types of MT output evaluation, for instance sentence-level in comparison to document-level, will also be discussed with one final argument on whether post-editing is easier and faster compared to translation from scratch.

In order to analyze the main difference between light and full post-editing in tangible examples, we need to understand the function this practice serves in the industry. Therefore, the conceptual dilemma between quality and speed needs to be addressed. After all, the reason why certain content is light post-edited depends entirely on the financial resources and the shelf-life of the content, making quality a secondary priority. When a certain content is destined to have a shorter shelf-life and, hence, the budget allocated is tight, customers tend to ask for light post-editing, not full. So, what is the difference between the two? Referring from Giselle de Almeida’s PhD thesis (2013), Moorkens and O’Brien define “‘light’ (or ‘fast’) post-editing” as the process where the aim is to “to create output that is understandable and readable, but not necessarily stylistically perfect”, and “‘full’ (or ‘conventional’) post-editing” as “producing quality equivalent to ‘that of a text that has been translated by a human linguist’” (2017, 2). Quality in both these definitions is clearly indicate

a linguistic one and excludes the interpersonal and communicative aspects of the translation process. Apart from the financial resources mentioned above, there is another concept of translation ‘quality’ which clearly points as both the process and product; that is, the “social quality”. As Moorkens et al. briefly state

Abdallah (2017) suggests that an all-encompassing quality model should include not only process and product, but also social quality, as the interactions and work practices of those in a translation network are likely to affect process and product quality (Moorkens et al. 2018, 2)

Yet, having given enough space to social constructs behind the process in the previous chapter, this chapter delves deeply into linguistic evaluation and the editing threshold between the two post-editing types. But the survey conducted before the evaluation process of this study is designed to have questions that address the issue of ‘social quality.’ Therefore, returning to the gigantic topic of evaluations, there are clearly overlapping problems, but also distinctly sharp points of view. The first issue to examine here is the development of evaluation metrics used by both the industry and the trainers when the previous translation criticism methods used especially by translation student trainers were considered lacking in many aspects.

3.1 From “optimal equivalence” to MQM

Historically speaking, there have been several translation quality assessment typologies used for pedagogical purposes in academic settings and also for the recruitment of freelance translators by the industry and so, their work was subjected to evaluation. In addition, with the introduction of the concept of ‘fitness for purpose’ – a concept that delineates that raw or post-edited MT is considered ‘good enough’ for “a notional translation end-user” – in translation quality assessment (TQA), crowdsourced evaluation became another topic discussed in the field. Yet, what is clear from the bigger picture is that of a “cross-pollination from one translation method to another” (Moorkens et al. 2018, 2), meaning what was a

trustworthy method of evaluation only for the industry came to be used in the pedagogical sphere as well, while the linguistics-based assessment methods used by the trainers continue to be the basis of quality assessment metrics used in the industry. ‘Fitness for purpose’ might be a vague and “relatively undefined” term. However, when we consider that it was born out of prescriptive guidelines for both types of post-editing and “the exacting requirements of detailed error typology evaluations with an elaborate system of associated penalties, or a combination of approaches tailored for a translation client” (Moorkens et al. 2018, 3), more empirical research is essential for various scenarios³⁰ and language pairs. But for the moment the discussion here focuses on the development of translation quality assessment metrics.

Castilho et al. (2018) starts the quality assessment discussion by expressing that theoretical discussion has common points in “the dichotomy between the source-oriented concept of accuracy (or adequacy) and the target-oriented concept of fluency” but in practice TQA processes vary by a fair amount and are subject to limitations particularly due to resource constraints. Also, the ubiquitous adoption of MT output in CAT tools has caused translation quality to be diversely interpreted and operationalized (10). For instance, referring to Doherty et al.’s research on the use of human and automatic evaluation, and also custom in-house measures alone or applied to their work, most participants said that they preferred human TQA (69%) whereas only 22% said they use automatic evaluation with 13% preferring customized in-house measures (10). In a similar light, it is almost an established fact for an overwhelming number of researchers in the field that “there is no single objective way to measure quality” (Drugan 2013 cited in Castilho et al. 2018) but when it comes to most sectors of the industry it is clear that “TQA is broadly limited to the application of somewhat arbitrary ‘one-size-fits-all’ error typology models that aim to give quantitative

³⁰ Moorkens et al. highlights “a new level of pragmatism” here, referring to usability of translation in a certain setting where “a sharpened focus on a targeted end-user (as detailed in Suojanen et al. 2015) has added a new meticulously calibrated variability to translation quality requirements” (2018, 3). Usability and quality evaluation require additional metrics to be designed and the translation brief to be provided to the translator. How they all merge in an automated translation project management system poses a question for further research. Nevertheless, usability is also addressed under certain branches of MQM and within the scope of this chapter MQM will be analyzed in detail. But no other usability metric will be discussed.

indicators of quality” (Lommel et al. 2014 cited in Castilho et al. 2018, 11). The underlying reason behind this ‘variability’ can be explained with respect to the function the evaluation process serves in industry and in research. In the former, the objective is “to ensure that a specified level of quality is identified, measured, and delivered to the client, buyer, end-user, etc., of translated content”, while in the latter the objective is generally “to obtain a measure that can show a demonstrable change in quality, most usually an improvement, from previous work or between different translation processes” (11). Thus, different objectives lead to different questions but in both settings the purpose is to reach an agreement on how to measure translation quality in an objective way (11) and on that note, a critical look into how the methodologies and approaches used for TQA emerged and converged, sometimes unified, is fundamental for both settings, especially when we consider different ways of translating including post-editing and interactive MT.

TQA metrics are designed based on theory, not on empirical data. Therefore, they present sometimes colliding and overly strict categories. The fact that “theorists disagree even on how many *categories* of models there are” is the evidence of research-oriented TQA being “heavily theoretically motivated, reflecting the assumptions, and to some extent the biases, of those who propose or adopt them” (Drugan 2013 cited in Castilho et al. 12). A number of researchers including House (1997), Schäffner (1997), Secara (2005), and Fields et al. (2014) among others have asserted that “evaluation is directly associated with the underlying translation theory that one subscribes to” and thus, alternating views on translation result in divergent concepts of translational quality and divergent ways of assessing it (House 1997 cited in Castilho et al. 12). But then, evaluation is not an issue based solely in the discipline but also in the industry. The initial answer to the question how the latter sees evaluation is understandably characterized by customer opinion. But, in fact, according to O’Brien (2012), QA in the industry is “managed by gatekeepers in the supply and demand chain who work with static evaluation models [. . .] applying penalties and maintaining thresholds with little, if any, input from customers”. Besides these criticisms of subjective and non-empirical usages, QA is integrated into CAT work environments, making comparisons between automatic and human evaluation inevitable. Yet, as Lommel states, the

two methods “have been fundamentally different in kind, preventing comparison of the two” (Lommel et al. 2014 cited in Castilho et al., 14). It is very common practice in the industry to carry out TQA as part of the process, sometimes concurrently with translation work, using either standalone or inbuilt CAT tool error checking software in terms of terminology, consistency, punctuation, formatting, number values, and tags³¹, or post-hoc using one of several evaluation models. Still, these tools “tend to generate a lot of false positives or incorrectly identified errors” (14) and language-dependent features of some give rise to more problems than solutions. Also, automatic evaluation or error checking tools generically assess the errors in a translation by comparing them to metrics which are “predetermined according to some logical or hierarchical criteria” and severity weightings predetermined by a senior translator or reviewer (14). There are currently different models with different parameters and error weightings in use presently such as LISA, ISO 11669, TAUS, QTLaunchPad and MQM. A detailed overview of these models is necessary for understanding the evolving nature of assessment from the early days with concepts such as “optimal equivalence”, and to the recently updated version of multidimensional quality metrics (MQM).

First developed and distributed in 1995, the Localisation Industry Standards Association (LISA) QA Model has probably been the most influential stand-alone tool for quality assessment in the industry, although the association closed its operations down in 2011. In the upgraded second version, which appeared in 1999, it had five major categories of errors, namely linguistic issues, physical issues, business and cultural issues and technical issues (LISA 2007: 12-14) and from the start, all the errors were categorized as “minor”, “major” or “critical” by the evaluator. Under linguistic issues, the model distinguishes seven types: mistranslation, accuracy, terminology, language, style, country and consistency. Each translated segment is assigned a score depending on the type of errors encountered and these individual scores lead to an overall score for the whole task. Depending on the threshold defined by the evaluator, a translation is given the status of “pass” or “fail”. The evaluator

³¹ Among the most widely used TQA tools are ApSIC XBench, Yamagata QA Distiller, ErrorSpy, Okapi CheckMate, and Acrocheck.

decides “how demanding and strict he/she wants to be in the overall evaluation of the translation product (i.e. how many errors of which severity can be tolerated in the evaluated text, for it not to be rejected)” (Castilho et al. 15). Thus, the major criticism aimed at the LISA evaluation was related to its “lack of empirical base” and the overlapping of some error categories (Jiménez-Crespo 2009 from Martínez Mateo 2014, 77). O’Brien’s criticism towards gatekeepers maintaining static evaluation models without any input from the customers is also pertinent in this context.

Another metric/guideline used at QA is ISO 11669 which “provides guidance concerning best practices for all phases of a translation project” (ISO/TS 11669:2012) and consists of 21 translation parameters in five categories, which are source content, requirements for the target, production tasks, environment, and relationships. The standard define translation quality as follows: “the quality of a translation – from a workflow and final delivery perspective – can be determined by the degree to which the target content adheres to the predetermined specifications”, when the two parties, both the service requesters and TSPs [Translation Service Providers] agree on project specifications (ISO/TS 11669:2012). Because of the earlier tendency to define quality in terms of text-specific features, the standard must have provided the field with an innovative but also empirical viewpoint on quality from a project management perspective. Defining translation quality from a text-centric perspective is not new as Castilho et al. (2018) points to these perspectives including Koby et al.’s (2014), and exclude services such as summarization, localization, etc. But an open-ended and vague definition also fails to serve the purpose since both parties involved in the process “do not always know what specifications a project requires, thus proposing a radically different view from that adopted in their broad definition” (16).

In that respect, another metric, DQF by TAUS, aims to fill “the existing gap between theoretical sophistication and the applicability” (Martínez Mateo 2014). The Translation Automation User Society (TAUS)³² Dynamic Quality Framework (DQF) was first developed

³² Martínez Mateo (2014) compares the text-centric metrics, e.g., LISA QA, SAE J 2450, TAUS Benchmarking and Quality Assessment Tool developed by Directorate-General for Translation (DGT), in

in 2011 “in close cooperation with many of the TAUS member companies” and its objective was “to standardize the methods and tools of quality evaluation, aggregate the scores and measurements and make these available through industry-shared metrics” (Valli 2015, 127). The major difference it presents is its consideration of variables such as communicative function, end-user requirements, context, mode of translation, including HT, raw MT and post-edited MT, and also content profiling and quality estimation as originating parameters to quality assurance. Another major difference it presents is that it approaches the problems before the translation process begins, not after the process is completed. In 2014, as part of QTLaunchPad Project, the German Research Center for Artificial Intelligence (DFKI) published the MQM (Multidimensional Quality Metrics) which focuses on building a customizable task-specific translation metric that could answer for various assessment needs. It provided a master vocabulary of error types. However, later, as a result of another European funded project, QT21, TAUS and DFKI harmonized their metrics, and the DQF and MQM error typologies were aligned into a single DQF-MQM framework in which high-level branches of errors match the six core DQF issue types. Castilho et al. further states that MQM was “developed to address the widely recognised limitations in previous quality evaluation models adopted in the translation industry, providing the flexibility to accommodate other standards”, referring to principles it has in common with LISA QA model (16-17). The trend these metrics evolve towards demonstrates an obvious effort to “reconcile, different views of translation quality, including, crucially, the expectations of end- users” (17).

At this point, it is time to analyze the seven high-level error type dimensions of MQM, which are terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, design and markup. LISA QA is very similar in terms of higher-level categories of MQM, whereas SAE J2450 QA³³, which this study does not analyze in detail, resembles the linguistic categorizations of translation criticism with the following

detail and argues that “they do not duly pay attention to the contextual (Sager 1989), the pragmatic (Nord 1997) nor the text-level issues” (88).

³³ SAE J 2450 was one of the pioneering evaluation frameworks/metrics for quality measurement on language translation specifically designed to be used in the automotive industry. For the determined scope and objectives, see “SAE” and Martínez Mateo (2014).

categories ordered according to their seriousness: wrong term, syntactic error, omission, word structure or agreement error, misspelling, punctuation error, miscellaneous error. On the other hand, the LISA categories echo TAUS DQF with higher-level branches such as doc language, doc formatting³⁴, help formatting, software formatting, and software functionality testing. The category doc language includes lower-level elements such as mistranslation, accuracy, terminology, language, style, country, and consistency, while doc formatting entails layout, typography, graphics, call outs and captions, TOC, and index. Categories related to software design involve the sub-categories of graphics, alignment, sizing, truncation/overlap, character formatting; localizable text, dialog functionality, menu functionality, hotkeys/accelerators, jumps/links. Overall, there are some very striking inevitable, shared parameters that all metrics try to measure. The first category is good old adequacy and fluency. Generally speaking, adequacy “is typically defined as the extent to which the translation transfers the meaning of the source-language unit into the target” (18) and it corresponds to the category of “accuracy” in MQM framework where this category is defined as “errors occurring when the target text does not accurately correspond to the propositional content of the source text, introduced by distorting, omitting, or adding to the message” (“themqm.org”). The adequacy judgement accompanies the comprehension of fluency, which concentrates on “errors related to the linguistic well-formedness of the text, including problems with grammaticality, spelling, punctuation, and mechanical correctness” (“themqm.org”), in other words, the target text and the degree the translation follows the rules and norms of the target-language. It is a noteworthy point that in the most recent version of MQM, is the higher-level category with this name was changed to “linguistic conventions”, further signaling its focus on the translation output. Some common error types in this category are incorrect pronouns, inconsistent prepositions, and incorrect punctuation. These two categories are commonly assessed using ordinal scales in the form of Likert scales, graded in the range between the lowest score of 0 to the highest one of 5. As a result, similar to the nature of editing and reviewing practices described in Chapter 2, assessing adequacy calls for some degree of bilingual proficiency, whereas

³⁴ Help formatting and doc formatting for Asian languages are distinct categories in LISA QA.

fluency requires, in principle, only proficiency in the target-language. Thus, experience in evaluation is not always considered a necessity in TQA and both professional and amateur evaluators can be employed to do the job (18).

The second category which acquired new contexts in the age of machine translation and large language models is readability and comprehensibility. Apparently, readability measures the ease with which a certain text can be read by one or more person(s) and is definitely associated with the suitability of a text for intended readers of a certain age and/or level of education. The means of measurement are typically linguistic such as word frequency and sentence length, as well as extra-linguistic features such as formatting and spacing. This group of parameters are reflected in the category “audience appropriateness” in MQM, which used to be called “verity” in version 1, defined as “errors arising from the use of content in the translation product that is invalid or inappropriate for the target locale or target audience” (“themqm.org”). Castilho et al. refers to Doherty’s (2012) identification of inconsistent usage in the literature “whereby comprehensibility and readability are either grouped together, where one is subsumed by the other, or where both concepts and their measurements are separate” (19). In MQM’s first version, the “subtypes” of the parent category of verity are completeness, including incomplete lists and incomplete procedure, culture-specific reference, end-user suitability, legal requirements and locale-specific content. Clearly, these subtypes suggest that information relayed through translation also satisfies the specific requirements of both end-user and market, providing an additional filter based on research on target locale conventions.

The third category deals with to what degree and according to what principles of production content is considered “acceptable”. Most research³⁵ on the acceptability of MT content interprets this parameter as “composed of various categories” such as usability, satisfaction, and quality. More precisely, acceptability is a quality attributed to output when

³⁵ Castilho (2016), who measures the acceptability of instructional content and also, Castilho and O’Brien (2016), Klerke et al. (2015), Doherty and O’Brien (2014), Castilho et al. (2014), Stymne et al. (2012), Roturier (2006). Nielsen (1993), as well as De Beaugrande and Dressler (1981). Yet, assessing the efficiency and usability of natural language processing applications through user satisfaction is more common in literature (e.g., Dybkjær et al. 2004).

users of that content³⁶ can use it ‘to perform tasks’ even though the content contains linguistic and other types of flaws, but as Castilho et al. (2018) puts it, “they will find the text less acceptable if the flaws in the translation affect their ability to use the text to some extent” (Castilho 2016 cited in 2018, 20-21). Therefore, it is of pivotal importance to understand what “usability” denotes, which ISO/TR 16982 defines as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” (2002), for the purposes of identifying the “inconsistency in terminology and operationalisation” (Doherty and O’Brien 2014). But it must be noted again that the standard focuses on usability as a generic concept in user-centred design in the fields, for example, of software engineering and the construction of interactive systems. In other words, usability of technical documentation is not the top priority. Following this overview of inconsistent usages of the evaluation metric, it is time to discuss how, meaning in what contexts and with which groups of people, human evaluation is performed in the overwhelming majority of studies.

The three significant lines of research when it comes to commonly recognized gold standards of research on MT output evaluation are comparing evaluation results of amateur and experienced evaluators, those of automatic and human evaluation, and finally of text- and document-level evaluation. The first one, because human evaluation is inevitably “resource intensive”, can take longer compared to human evaluation of human translations and can be biased due to a number of factors such as use of “forced-choice dichotomies of good versus bad translation” and interval scales, as well as subjectivity; hence, most research outside TS tends to rely on automatic evaluation where an iterative system is embedded in the CAT tool or outsourced software including simple spreadsheets which enable the user to produce a quick evaluation result (Doherty 2017, 3-4). As a consequence of these resource constraints, it is hard to access trained professional evaluators, hence the tendency to rely on students and amateur evaluators whose proficiency in the languages involved is not defined or whose expertise level is unknown and who may not necessarily have any training on TQA or error taxonomies. Therefore, the practice of having a group-based and crowd-based TQA

was invented to compensate for the lower level of expertise the amateurs may display when asked to elicit their judgements on quality. In the opposite case, it is very unusual to find groups of professionals or experts assessing MT output due to the mounting costs. Ensuing from this common practice, data available from professional or trained evaluators has proved to be an exception in evaluation tasks, which further renders the reliability of evaluations problematic, and inter-rater agreement thorny (9).

An alternative to this ‘end-of-the-line’ problem, which only focuses on the end product and entails that “any changes can be costly both financially and in terms of missed deadlines” was offered by Suojanen et al. (2015); and this mainly academic model is called user-centred translation (UCT). As companies are trying to step beyond a rush-to-the-bottom, this new model suggests ways of becoming “more versatile and more innovative, offering translation-related services that are clearly identifiable and bring added value to the customers” (2) for it “concentrates on imagining what kind of a process will produce a variety of successful translations to serve the needs of different commissions”, hence the concept usability evaluation (128). On that note, the authors add “what is at the heart of evaluation and of potential corrective procedures are those weaknesses that are critical or significant to relevance, and not just any differences in equivalence between the source text and the translation” (133). Yet, in terms of assessing usability, this is not simple and free. The authors refer to Nielsen’s classification of usability problems which presents a customized Likert scale with 0 meaning no problem, 1 a cosmetic problem only, and 2 minor, 3 major and 4 meaning a usability catastrophe (132).

Efforts to evaluate MT output are not new at all. Back in the MT Summit IV in 1993, Yorick Wilks famously said and wrote “MT evaluation is, for all its faults, probably in better shape than MT itself” (139). The history of building a methodology for evaluating MT systems, i.e. an automatic evaluation system, dates back to 1990s. The ultimate goal of one³⁷ of these automatic metrics was to “build a coherent picture of the various features and metrics that have been used in the past, to offer a common descriptive framework and vocabulary,

³⁷ Among these notable MT and automatic metric designing projects were The Defense Advanced Research Projects Agency (DARPA) MTE and Framework for the Evaluation of MT (FEMTI).

and to unify the process of evaluation design” (Hovy et al. 2002). They also present additional advantages in terms of budget limitations, speed compared to manual evaluation and the supposedly unbiased results of evaluation. Yet, the results will inevitably be less comprehensive than manual evaluation and do not “readily indicate the type of problems that the translated text contains”, as they cannot go beyond “mere adequacy and fluency” (Castilho, Doherty et al. 2018, 25). However, in that case, research demonstrates that manual evaluation has its downsides, primarily inter-annotator agreement. Having multiple judges is desirable on paper but they do not always “have a good grounding in translation and linguistics” and the results might be misleading due to the varying skill levels of the evaluators. For instance, for judging grammaticality and fluency looking at the target text alone is enough but judgements on accuracy and adequacy require bilingual competence, which some of the evaluators in previous research did not have. Therefore, the error categories assigned to errors end up being faulty.

The mechanism in which these metrics work explains the advantages and disadvantages they have compared to against human evaluation. Koehn (2010) states that the main purpose of automatic metrics is to compare the MT output, or as in some studies “translation hypothesis”, to one or several reference translations which are considered accurate and of ‘good quality’ simply because they are human translations³⁸ and compute a similarity score to them. Though there are several new metrics, the most common ones are BLEU, NIST (National Institute of Standards and Technology), METEOR proposed by Banerjee and Lavie (2005), TER (Translation Edit Rate) and chrF proposed by Popović (2015). The first one, BLEU (Bi-Lingual Evaluation Understudy) is considered to be the de facto standard used for comparing various MT engines’ output and is also used in comparison with METEOR and TER for most research purposes. Most of these metrics work by using *n*-grams as the parameter for measuring. Way (2018) opines that there are some inherent problems with automatic metrics, the major one of which is that all of them

³⁸ Yet, in recent discussions in WMT the question whether reference translations are definitely produced by humans (professional translators or not) has been raised and the results of the shared tasks have been questioned for that reason. The gold standard quality of these reference translations is an assumption and not verified (Castilho, Doherty et al. 2018, 26).

ignore the source sentence altogether. Another problem Way points out is that these metrics produce better results when they measure similarity of output in comparison to multiple human-supplied references (2018, 167). But most of the time, only a single reference is supplied. Thus, the assumption that the reference is the one and only correct translation of the ST will cause unreliable results, and the studies in the next part demonstrates that ‘essential’ changes (language errors and mistranslations) comprise only about half of the overall edits made on some texts (de Almeida 2013 cited in Way). For this reason, Castilho, Doherty et al. (2018) suggests that researchers are concerned as to the actual value of automatic metrics (26). Way also asserts that “certain automatic metrics prefer shorter/longer translations” (2018, 167). While METEOR gives higher results for “longer outputs than the reference translations owing to the different chunk penalties assigned to different languages” (He and Way 2009a, b), TER prefers shorter sentences. Additionally, because TER is an error metric, as an MT system generates more words, a higher the number of insertions, deletions and substitutions will be required to transform the output into the reference sentence. He and Way (2009a) note that this does not harm the evaluation but “if such knowledge is made use of in tuning, the system will be *tuned* to take advantage of this preference, and will tend to output overly succinct sentences” (Way 2018, 167).

For all these reasons, automatic MT evaluation and evaluation estimation³⁹, just as MT itself, face challenges in assessing syntactic and semantic equivalence. One final development, in this respect, has been achieved with evaluation at the document level rather than sentence/segment level. Using measurements based on coherence and discourse structures for document level evaluation has been one of the methods tried. For instance, including reference and meta-evaluation methods in their mechanism (Giménez et al. 2010), while using cohesion devices/ content words (Wong and Kit 2012) and incorporating discourse structures besides existing automatic MTE metrics (Guzmán et al. 2014) were among the other attempts in the field. Yet, it is clear automatic evaluation will be used

³⁹ Measuring and means of measuring have been focal points in research where various experiments try to incorporate comparisons of multiple bits of texts to take the level of linguistic analysis and understanding of AI one step further.

alongside human evaluation for some time in the future and therefore, the questions Doherty addresses are still relevant at present:

[W]ill students and researchers from TS be biased in evaluating MT quality? And will MT researchers also be biased in the same evaluation? Coupled with this is the reliance on self-selective sampling and self-description of participants who are not always the best judge of their own language proficiency and in the case of online crowdsourcing may not be truthful in describing their expertise and professional experience (Doherty 2017, 11).

In the light of this discussion, assessing the MT output quality and various uses of that output lead the evaluation schemes to have two divergent categories, i.e. light and full post-editing (PE), which require disparate levels of effort but in either case, a specific quality level is achieved. Light PE is encouraged where the content has a short shelf-life and so, the post-editors are asked only to correct essential errors and hence, the process of post-editing is, in theory, much faster. In contrast, full PE requires more corrections for a higher quality in contexts where the shelf-life of the content is assumed to be longer, or the content will be in use where it is highly visible or has sensitive information (O'Brien et al. 2009) but types of PE are used to achieve publishable quality for the final revised MT output. In a pioneering work, Krings (2001) divided the PE effort into three categories: temporal, technical, and cognitive, where temporal is defined as the amount of time spent doing PE, technical effort as the number of edits made by the post-editor and cognitive is measured using methods such as eye-tracking, keyboard logging, pause-to-word ratios. Numerous studies were done to measure effort required for both levels of PE (g. De Almeida and O'Brien 2010; Depraetere 2010; Plitt and Masselot 2010; Sousa et al. 2011; Specia 2011; Koponen 2012; O'Brien et al. 2012, 2013, 2014; Guerberof 2014; Lacruz and Shreve 2014; Carl et al. 2015; Daems et al. 2015; Moorkens et al. 2015), with questions asking whether PE would be time- and cost-effective compared to translation from scratch and also, whether students and professional translators already in the market would be able to do the job. The underlying purpose of the first question is to pre-determine pricing decisions and turnaround time, making certain markets more competitive where budget-conscious clients see nothing untoward in asking

translators to meet tight deadlines but not compromise on the final quality of the translation product. Therefore, reliably assessing the expected PE effort is indispensable as it enables LSPs to optimize their translation processes, hence the question how to quantify the amount of effort that is necessary for MT PE (29).

A number of fundamental issues arise here in terms of research, education, and professional practice. Castilho, Doherty et al. (2018) lists them as the following: lack of standardisation in TQA usage, inherent inconsistency in TQA, the relationship between human and automatic measures, social quality and risk, education and training. At the crossroads of buyers and suppliers, what quality denotes is a question of much dispute. As Ray et al. (2013) state, there is a significant lack of agreement about what QA means and how it is measured, which inevitably results in incompatible expectations and unsuccessful projects. The authors underline that there are “no generally accepted standards” in TQA, where “quality expectations tend to follow prices up but not down” (ibid.). In the case of research on QA, similar issues persist specifically in the recruitment of inappropriate evaluators and the lack of transparency in specifications and procedures and also relying on one type of automatic evaluation over a holistic project-based approach which would (ideally) minimally reflect issues inherent in manual or automatic evaluation, in other words “a lack of standardisation at the macro and micro level of TQA” (Castilho et al. 2018, 30). There has been a set of guidelines proposed in order to limit subjective judgments and arbitration, and also, QA “boiling down to the opinion of one person over that of another” with “too much time in arbitration mode” in DePalma and Kelly’s words (2009). For instance, one set of guidelines proposed by Doherty (2017) attempts to improve validity, reliability, and consistency by adopting psychometric principles from the psychological and cognitive sciences and also stresses evaluator training and materials. Not surprisingly, the results of a number of studies demonstrate “considerable variation in the correlations between human and machine TQA (see, e.g. Labaka et al. 2014; Doherty 2017)” and thus, quite often “the results of manual and automatic evaluations do not agree” (Labaka et al. 2014 cited in Castilho et al. 2018, 31). The main problematic areas that arise in the face of evaluation will increase when we multiply the evaluation parameters with different evaluator groups with

divergent levels of translation experience and attempt to categorize their actions in the four possible segments for dealing with MT output, i.e., accept without any changes, light post-editing, full post-editing and reject the output, hence translation from scratch. What's more, the effort spent on each task undoubtedly varies in each evaluator group, as well as their translation behavior, including all the actions deemed necessary for research. In the next part, empirical studies on post-editing behavior and the effort associated with them are analyzed in detail.

3.2 Requirements for full and light post-editing according to ISO 18587

Before analyzing the results of the studies on post-editing behavior and effort, it is crucial to understand the difference between two types of PE explained very briefly in the international standard ISO 18587. In that case, the standard defines 'post-editing' as "edit and correct machine translation output", turning a blind eye to the issue of editing translations found in the memories. The first type of PE, i.e., "full post-editing" is defined as "process of post-editing to obtain a product comparable to a product obtained by human translation" (2017, 2). The aim of this level of PE is to produce an output which is "indistinguishable from human translation output" (8). But the clause in the standard continues with this sentence: "Nevertheless, it is recommended that post-editors use as much of the MT output as possible" (8). The standard advises post-editors assigned with this job to focus on the following categories:

- a) ensuring that no information has been added or omitted; b) editing any inappropriate content; c) restructuring sentences in the case of incorrect or unclear meaning; d) producing grammatically, syntactically and semantically correct target language content; e) adhering to client and/or domain terminology; f) applying spelling, punctuation and hyphenation rules; g) ensuring that the style appropriate for the text type is used and that stylistic guidelines provided by the client are observed; h) applying formatting rules (2017, 8)

The end product of full PE corresponds to that achieved after the revision process in the (human) translation workflow. The post-editor is also asked to analyze the stylistic guidelines and apply the necessary rules to the product. Terminological consistency is achieved, and most importantly, if certain parts of the ST are missing in the MT output, then the post-editor is asked to complete translation, therefore presumably to translate from scratch. This final point makes the overall distinction between translating from scratch and full post-edit as the amount of missing information can simply change due to the MT paradigm used. Though, NMT and LLM⁴⁰ paradigms are known not to miss big chunks of ST in translation, the amount of translation from scratch that will be necessary also depends on the domain type and the amount of resource for a certain language combination. For this reason, the pricing of full post-editing is an issue much discussed in the industry, open to negotiation from all parties involved. This brings us to ask what difference there is between full and light post-editing. ISO 18587 defines ‘light post-editing’ as “process of *post-editing* to obtain a merely comprehensible text without any attempt to produce a product comparable to a product obtained by *human translation*” (2017, 2). Therefore, this level of post-editing requires the post-editor to have a high-quality MT output, which in turn is a result of high resource language combination and domain. As the objectives of both light and full post-editing are “comprehensibility of the post-edited MT product” and also, “correspondence of source language content and target language content”, as well as the usual “compliance with post-editing requirements and specifications defined by the TSP” (6). Surprisingly though, the main body of the international standard concentrates on full post-editing, while light post-editing is only more intricately examined in the annexes because, as the document states, this level of post-editing is “normally used when the final text is not intended for publication and is mainly needed for information gisting i.e. for rendering the main idea or point of the text” (10). On this point, the standard warns that this process does not “need to” produce a “stylistically adequate” translation, yet it shall be “comprehensible and accurate”. Moreover, the document lists the points post-editors should focus on:

⁴⁰ For the most recent book on the topic of LLMs comparing LLMs and MLLMs in terms of good performance both in quality and scope, see Moorkens, Way and Lankford (2025).

- a) using as much of the raw MT output as possible; b) ensuring that no information has been added or omitted; c) editing any inappropriate content; d) restructuring sentences in the case of incorrect or unclear meaning.” (2017, 10)

Here, the definition focusing on gisting purposes only is not exactly in line with the requirements. When compared with the requirements of full post-editing, the post-editor is not asked to stick to the stylistic guidelines of the client or ensure stylistic consistency in the light post-editing process. But apart from the fact that (s)he is asked to use as much raw MT output as possible, requirements for light post-editing complicate the differentiation between translation from scratch using the already available translation memories stored in a CAT tool, revising and full post-editing. The question we encountered in the previous chapter arises again: Does full post-editing replace revision in the translation workflow where human translation is replaced by MT output? The easiest answer to this question can be found in the real practice of professionals and, how newcomers to the industry, i.e., novice translators, interpret these requirements. In the next part, we take a deeper look into the discussion on the two levels of post-editing and effort associated with them.

3.3 Light post-editing and full post-editing: The discrepancy between ISO in terms of definitions and what different experience groups deem necessary for each level of post-editing

One of the earliest classifications of post-editing was written by Jeffrey Allen in an article published in the renowned book edited by Harold Somers with the title *Computers and Translation: A Translator's Guide*, where he differentiates between an inbound and outbound translation approach. An inbound translation approach is explained as “the process of translating to understand”, whereas an outbound approach is described as “the process of translating to communicate” (2003, 301). The level of post-editing is dependent on several factors such as “the expectation with regard to the level of quality for reading the final draft of the translated product”, “the use of the document with regard to the life expectancy and

perishability of information” and “the use of the final text in the range from information gisting to publishable information” among similar others (ibid.). Under inbound approaches, Allen lists MT with no post-editing and rapid post-editing, as both categories imply these approaches are intended only for information purposes and the MT output does not go through any editing and/or revision process in them. On the other hand, under outbound approaches Allen lists “MT with no post-editing”, “minimal post-editing” and finally, “full post-editing”. The no post-editing approach proved to be “limited in use and applicability for information dissemination activities” (304). Therefore, similarly to the two categories ISO 18587 offers, the final two approaches and concepts were largely adopted. Here lies the need for empirical studies into setting the threshold between the two types of post-editing as Allen defines the concept of *minimal* PE as “a fuzzy, wide-range category because it often depends on how the post-editors define and implement the ‘minimum’ amount of changes to make in view of the client/reader audience” (ibid.) and continues “the level of sensibility of interpreting the concept [...] often seems to vary from one post-editor to another, from one manager to another, from one reviser to another” (305), while he explains “[t]he idea of full post-editing of texts has been debated for many years because full post-editing implies a high level of quality of the resulting texts” (306). Dozens of studies on the topic has been carried out in the two decades since Allen wrote those lines but the general understanding is that “the aim of PE is to achieve publishable quality for the final revised MT output”, thus it requires “more corrections for a higher quality, with a slower turn-around, and is typically employed to disseminate high-visibility or sensitive texts” (O’Brien et al. 2009 cited in Castilho et al. 2018, 27). As in the international standard, minimal PE was renamed ‘light PE’ and requires only essential errors to be corrected. The crucial point here is who provides a more accurate evaluation of content in terms of which evaluation framework and how this evaluation aligns with the requirements of the two PE approaches. For answering this detailed and complicated question with more than one interrelation, this study suggests the literature review below provided under this sub-heading.

On that note, we start with Arle Lommel’s arguments and evidence for standardising error typologies, i.e., MQM, TAUS DQF and the harmonisation of the two within the EU-

funded QT21 project in 2014 and 2015 (2018). Lommel states that MQM and DQF began separately but they were perceived to be competing projects and later on, with the QT21 project, they were successfully harmonised to design an emerging de facto standard for TQA (110). Before standardized quality assessments, the best practice in the language industry was having a bilingual reviewer or a monolingual subject matter expert review translations and give a final rather informal evaluation. These evaluators did not typically use “formal, predefined rubrics or tools for evaluation”, hence there was no common best practice in the industry, which proved to be a major inconsistency that caused “a source of confusion for clients” (110). There were two sides of this to assessment coin: As LSPs faced pressure from their clients to assure the quality of translations were adequate, translators perceived them to be tools to penalise them or renegotiate prices (111). Similar to the TQA tool embedded in CAT tools, the earlier method for assessing quality was typically in spreadsheet form in which reviewers counted errors for each category but they could also assign them weights from minor, through major to severe in order to assign different penalties. All in all, the error categorizations remained “ad hoc” and varied from LSP to LSP and the errors spotted in the translations were not linked to specific locations in the text. These earlier approaches to QA resulted in two problems, as Lommel puts it, “(i) the scores were ultimately unverifiable because the only link to the text was in the mind of the reviewer; (ii) it was unclear if the scores they generated correlated with audience or customer requirements” (111). In terms of the concepts we use today to assess the evaluations on MT output, these approaches mean very low inter-annotator agreement (IAA) could be achieved because reviewers were not interchangeable. They could disagree about the error category and also the severity of the error, if it was even an error in the first place. Additionally, these models of quality assurance were only built with specific scenarios or text types in mind which did not suit different scenarios or text types. For all these reasons, Lommel writes, the LISA QA Model underwent numerous changes and was frequently updated (112). In the 1990s, two systematic approaches to ad hoc TQA were developed, namely SAE J2450 and LISA QA. The very first standardized metric was called SAE J2450 which openly addressed the problem of scope, and hence domain, by stating that it was intended only for automotive service manuals. LISA

QA, which we analyzed above, was followed by its successor, Globalization Metrics Exchange – Quality (GMX-Q), which was “intended to be a companion to the existing LISA standard GMX-V (for the volume of translation) and the planned GMX-C” (113). Yet, due to LISA’s closure in 2011, the release of this final metric was stopped.

Subsequent to the failed release of LISA’s GMX-Q, MQM took up the job of designing an up-to-date metric to be used for both human and machine translation by adopting principles such as a flexible catalogue of error types, compatibility with existing specifications and tools, a hierarchical approach, and a specifications-based approach (113). Considering the categories we analyzed in detail above, MQM was designed to provide evaluators with a master vocabulary that could be specified to include minute details. It does not “define a single metric”, yet much like the TermBase eXchange (TBX, ISO 30042:2008) standard that presents an XML vocabulary for terminology databases, it serves as a common vocabulary for any metric in any domain and language (113).

Another key quality of MQM is that it is the thorough outcome of harmonization efforts comprising nine error typologies and tools already in use, i.e., the LISA QA Model, SAE J2450 and ISO CD 14080 metrics, error categories of SDL Trados Classic, ApSIC Xbench, Okapi CheckMate, the XLIFF:doc specification, and Yamagata QA Distiller, and finally, the grading rubric for ATA certification examinations (114). All these metrics had 145⁴¹ error types in total, with the most common error types being terminology/glossary adherence, omission, punctuation, with consistency, grammar/syntax and spelling specified equal times in order (114). Because it was created to be a superset of QA tools, “extremely detailed issues” and issues related to project satisfaction from the LISA model were omitted. For instance, instead of using fine-grained distinctions, it covered 12 issues related to white space within translation segments under a single category called “whitespace”.

Probably the most characteristic quality of MQM is its highly hierarchical structure. For the most specific ‘child’ categories the tree format extends up to four layers with increasing specificity but mostly stops at the second or third layer (116). Herein lies a

⁴¹ Lommel states that the initial drafts of MQM had 104 issue types, which increased to 182 in the final project.

dilemma for evaluation. Because as course-grained as possible, though nonetheless still serving the purpose, some categories are often too confusing to distinguish between, which causes evaluators not always to agree on particularly lower-level sub-issues, even though they generally do so on high-level categories. On this issue, Lommel advises that [i]n cases of uncertainty, parent categories should be used because they represent the point of least uncertainty and their children also count as examples of their parents” (116-17). The table below demonstrates the number of issues under each high-level ‘parent’ category:

Table 2: MQM Full Error Categories updated in 2024

Terminology	Accuracy	Linguistic conventions	Style	Locale conventions	Audience Appropriateness	Design and Markup
- Inconsistent with terminology resource (2) - Inconsistent use of terminology (3s) - Wrong term (1 sub-issue)	- Mistranslation (10s) - Overtranslation - Undertranslation - Addition - Omission (1s) - Unjustified euphemism - Do not translate - Untranslated (1s) - Retained factual error - Completeness (2s)	- Grammar (4s) - Punctuation (1s) - Whitespace - Spelling (4s) - Title style - Corpus conformance - Pattern problem - Duplication - Sorting - Unclear reference - Unintelligible - Character encoding - Non-allowed characters - Textual conventions (4s)	- Organization style(7 sub-issues) - Third-party style - Inconsistent with external reference - Language register (2 sub-issues) - Awkward style - Unidiomatic style - Inconsistent style	- Number - Currency - Measurement - Time - Date - Calendar - Name - Address - Postal - Telephone - Locale-specific punctuation (1s) - National language standard - Shortcut key	- Culture-specific reference (4s.) - Offensive (5 s.)	- Character formatting (2s) - Layout (6s) - Graphics and tables (3s) - Markup tag (5s) - Truncation/text expansion - Missing text - Link/cross-reference (2s)

With the highest number of issue and sub-issue types, the most common parent category in MQM, just as all metrics and frameworks is linguistic conventions⁴², defined as “errors related to the linguistic well-formedness of the content”, which answers the third question in the MQM decision tree. The evaluation starts with the terminology branch, which described “issues related to domain-specific terminology” (Lommel 2018, 118). If the evaluator does not find the error related to this branch, (s)he jumps to accuracy, which consists of “errors occurring when the target text does not accurately correspond to the propositional content of

⁴² Linguistic conventions was named “fluency” in MQM version 1.

the source text, introduced by distorting, omitting, or adding to the message” (MQM). One of the most striking change was made to Accuracy’s issue types in version 2 with the addition of “MT hallucination” under ‘Mistranslation’, which is defined as “[m]achine translation that is completely decoupled from the sense of the input sentence” (MQM Full). Style-related errors are those which “deviate from organizational style guides or exhibit inappropriate language style”, while locale conventions deal with errors violating locale-specific content or formatting requirements. Audience appropriateness is concerned with “[e]rrors arising from the use of target content that is invalid or inappropriate for the target locale or target audience”, while design and markup focuses on “[e]rrors related to the physical design or presentation of a translation product” (“MQM Full”). A big branch titled “internationalization”, with 49 issues in version 1, was replaced by a combination of issue types under the other branches. Another difference found in the second version is that of branch name change where Verity which was, in Lommel’s words, “a novel contribution from MQM to the broader study of translation quality” (2018a, 118), was renamed “audience appropriateness”, because verity refers to legal or regulatory texts and technical documentation, which “require substantial adaptation to correctly refer to laws and regulations in the environment in which they occur” (ibid.). Thus, the name of this branch was made clearer in the second version. If no further specificity is demanded from the evaluator, each branch can serve as an issue type on its own.

Lommel advocates that “MQM embraces a functionalist perspective at its core” though it is obviously the outcome of a standardising, hence universalising, effort (119). The two notions might seem contradictory since designing a set of criteria implies an assumption of the rules “*all translations should follow*”, whereas a functionalist or Skopos-oriented notion of translation stresses the idea that translations serve purposes and should be evaluated in those terms (Nord 1997). But when the branches are considered thoroughly, the reader can easily observe that MQM criteria are definitely function-based, stressing the qualities/conventions of the translated work’s environment. In other words, the criteria measures how well a translation meets the specifications at the receiving end and therefore the target-language purpose of this content has the potential to be dramatically different from

that of the source-language. For this reason, MQM is designed to be customized according to a specifications template that follows ASTM F2575-14⁴³ from the requestors and, as a consequence, feedback can also be tied to requirements. Lommel notes that creating new specifications and metrics for every project would be “wasteful and confusing” and MQM does not expect such a practice. Rather, it is intended that implementers “create specification templates and default metrics for different project types and reuse them to promote clarity and consistency” (Lommel 2018, 120).

Another point about MQM is deeply-related to different levels of post-editing. It concerns severity attributed to each error and issue type, but Lommel very carefully differentiates between two “distinct concepts” which are severity and importance. Severity points to “the nature of the error itself and its effects on usability of the translation”, whereas importance is indicated through the use of *weights*. While severity focuses on individual errors, not on entire categories, importance is concerned with, let’s say, the fact that style may not be as big a problem for technical documentation as it would be for marketing material (120-21). The four severity levels MQM supports by default are: critical, major, minor and null. Evidently, critical errors are those which render a project unfit for purpose on their own, and even a single error would hinder a project in fulfilling its purpose and hence, would cause trouble for project specifications; for instance, inaccurate information in terms of safety or legal implications. Major errors cause the intended meaning of the text to be unclear such that “the intended user cannot recover the meaning from the text” but are not likely to cause any harm to the translation user (120). Minor errors, on the other hand, do not have an impact on usability of the content and most of the time will be corrected by the intended user or the user may not even notice them. The null level was added to MQM in 2014 to enhance its compatibility with the TAUS DQF and is used to spot changes that are not considered errors. Implementers are free to assign weights to specific issue types according to their project specifications with the default weight being 1.0. How, then, is a

⁴³ ASTM F2575-14 is Standard Guide for Quality Assurance in Translation which serves all parties in the translation environment.

score calculated? Each error is multiplied by its severity value and its weight to generate penalty points and summed up in the point total. Finally, the score is calculated using the following formula: $\text{Score} = 1 - \text{Penalties} / \text{Word Count}$ and is generally presented as a percentage (121-22). If the implementer is aiming for a score for Pass/Fail, Acceptable/Unacceptable etc., then they must set a threshold.

One final quality of MQM is its capacity to serve both holistic and analytic translation evaluation. The difference between these two approaches is a clear-cut one. Holistic evaluation serves the purpose of obtaining a big picture image of translation quality, and therefore looks at the project as a whole, whereas analytic evaluation is used for “identifying and documenting specific problems, but may not capture the overall impression” (122). Needless to say, analytic evaluation is resource-intensive and requires training to apply. Determining what issues – generally high-level issues – the evaluation metric will contain and using a Likert scale to assess the issues can yield two results. Either the holistic evaluation gives an adequate score and does not indicate any problems, in which case there is no need to do a detailed analytic evaluation, or the translation fails a holistic evaluation and, in this case, Lommel states that there is “no need to invest the time in a detailed analytic evaluation because the reviewer already knows it would not pass” (Lommel 2018, 123). Given that MQM and DQF error typology were designed by two separate organizations, there are broader-scale differences between the two frameworks. The latter, DQF Error Typology, was designed to answer the needs of TAUS’ localization-oriented members and does not address all the possible error issue types that MQM covers but focuses on solving specific areas. It has six⁴⁴ error types, including accuracy, linguistic, terminology, style, country standards and layout. The DQF contains four severity levels which approximately correspond with the MQM severities. As a result of the integration⁴⁵ efforts of the two organizations,

⁴⁴ TAUS DQF also features four additional categories which are query implementation, client edit, repeat and kudos.

⁴⁵ MQM was developed at the German Research Centre for Artificial Intelligence (DFKI) by a group of researchers from 2012 through 2014 as part of the European Union-funded QTLaunchPad project, while TAUS Dynamic Quality Framework (DQF) Error Typology was developed by the Amsterdam-based

MQM version 1 was released in 2014 and was re-released as version 2 in 2024. The change in the MQM dimensions ensuing from the integration was not significant as an issue type: Terminology, under Accuracy and Fluency, was simply promoted to a full dimension (124-25). The major difference in DQF was its simplification by TAUS and the names of their categories were updated to comply with MQM (125). The severity of error types the present study uses is in line with MQM version 2.

Besides metrics and frameworks to measure quality, one parameter that differentiates light and full post-editing is people and their expertise. Similar to the test group of this study, the literature abounds in the assessments of the quality attributed to a specific translation and/or MT output with groups of people having divergent expertise and experiences. In this sense, Jiménez-Crespo's article (2018) on crowdsourcing and its impact on the fitness for purpose model in translation project management sheds light on the notion of quality. His argument starts from a well-discussed point in the industry, and also in the standards as he articulates, common approaches to governing procedures for achieving quality instead of referring to normative statements about what constitutes quality are the best practices of today. Therefore, these procedures are “generically process-oriented, rather than product-oriented” (76) and though not explicitly admitted by all, the constructs of these processes are somehow embedded in “either explicit theoretical approaches or implicit and a- or pre-theoretical ones held by individuals or collectives” (75). For instance, LSPs and their smaller companies, including Gengo, Getlocalization, Speaklike, and Unbabel offer distinct quality tiers matched with ‘best-suited’ content types such as ‘Basic’, ‘Marketing’, and ‘Specialised’ (77). Because they offer services as or for “self-serve translation platforms”, each level seems to entail “a different type of crowd, from bilinguals to vetted and/or ‘certified’ professional translators” (77). As a result, the notion of fitness for purpose emerges here, indicating an understanding of quality that is acceptable or adequate, which depends entirely on “the

Translation Automation User Society (TAUS). The harmonization of the two frameworks was executed as a collaborative effort by DFKI and TAUS within the EU-funded QT21 project in 2014 and 2015. No PE guideline had been published before 2010 and biggest LSPs had their in-house customized guidelines according to Flanagan and Christensen (2014, 260).

complex interaction of user attitudes, business decisions (or lack thereof), content types, translation types, and the relative importance of the translated text” (78). The author quotes at this point from Drugan (2013)’s earlier definition of this model as a “conscious attempt at using translation and revision resources intelligently” (79). In this sense, translation quality is “a time, resource, and financially constrained process” in this new paradigm (78), whereas classical approaches put responsibility for translation quality on “the shoulders of both LSPs or institutions and qualified translators and revisers” (79). LSPs were given the choice of complying with the ISO17100 requirements⁴⁶ and quality control procedures. Yet, in the new paid crowdsourcing paradigm responsibility has shifted towards the client or customer, who must decide which quality level matches their content or how much they value the localization/translation of their content. At this point, the participants in the pool are responsible for meeting the client’s quality expectations. But what happens when clients end up with low-quality, disfluent translations? The solution, to get high-quality work from non-experts, is to consolidate quality through “soliciting multiple translations, redundantly editing them, and then selecting the best of the bunch” according to Callison-Burch (2011). It must be noted, though, this new paradigm was specifically used for obtaining translations to feed and improve MT engines and the main focus was not on getting high-quality work (Zaidan and Callison-Burch 2011; Zbib et al. 2013; Gao et al. 2015; Yan et al. 2015). Connecting the crowdsourcing paradigm to user-based approaches from research by Nida (1964) and Nida and Taber (1969), Jiménez-Crespo describes this new paradigm as “a response-oriented or behavioural approach to translation evaluation” (86). The two practical tests to guarantee quality, i.e., eliciting the receiver’s reaction to several translation alternatives or reading aloud the translated work before an audience, are the closest evaluation methods to current iterative mechanisms (87). The many eyes principle, which ISO17100 explains in detail, originates from the German Federal Data Protection Act (BDSG) which came into force in

⁴⁶ ISO 17100 can be interpreted as an outcome of research on quality from “professional (Gouadec 2007), pedagogical (Kelly 2005), and cognitive approaches (Jääskeläinen 2016) to translation” (Jiménez-Crespo 2018, 80).

1997. But it was first formulated by Raymond (1999) as “Linus’ law”, referring to the practice of having more than one person in open-source software development and entails the well-known quote “given enough eyeballs, all bugs are shallow”. This principle assumes that given a large enough crowd perusing the same material, be it software code or translation, any existing flaw or errors will be identified and corrected (87). Previous research has investigated whether crowdsourcing could be used for translation evaluation in place of automatic MT evaluation measures such as BLEU (Papineni et al. 2002) or whether crowdsourcing results could be compared to actual professionals’ assessment of MT output for system comparison (Goto et al. 2014).

Three key issues stand out as main uses of crowdsourcing: one is “the consolidation of the process-based understanding of quality that permeates the language industry”, the second is “the push to further consolidate the fitness for purpose paradigm”, and the third is “the sharing of responsibilities in terms of quality among the various participants in translation efforts” in Jiménez-Crespo’s words (88). At this point, one major question is relevant again. What can be the new uses and trends upon which AI-driven business models and workflows exert an influence on in the language industry, specifically in terms of quality? After all, in the early years of MT post-editing the developers were very keen to use expressions such as “human parity” and “super human performance” (Poibeau 2022) and Doherty (2016) reports that the volume of translation produced using PE does not exceed 5%, even though PE was considered to have become the norm in the industry (i.e. Garcia 2010). Beyond the over-ambitious statements about this fuzzy notion of ‘quality in translation’, crowdsourcing would hardly elucidate the threshold for editing and quality, but undoubtedly, it has a place in the interrelation of MT, paraprofessional translation, and high-quality professional translation. For this reason, the present review here analyzes the end-user perception of content quality.

Gys-Walt van Egdom and Mark Pluymaekers’ article on text quality, text perception and the editing threshold sheds light on the interrelation of these topics with some unexpected results. Experimenting with the fitness for purpose paradigm, they report on a study where they asked end-users to pass judgement on manipulated machine translations with different

degrees of post-editing, using an informative and an instructive text. Participants graded the texts using a 7-point Likert scale on dimensions of text quality including content, language use, text logic and terminology, style, and finally, usability. In the traditional approach to quality, the authors elaborate, “quality tends to be equated with ‘excellence’ or ‘flawlessness’” but instead of sticking to the traditional approach, they attempt to “judge post-edited MT output on its fitness for purpose, which we conceptualize as the degree to which a text fulfils the end users’ informational needs and achieves its communication goals” (158-59). Despite the stress put on flawlessness, the industry has managed to find new methods and workflows such as crowdsourcing “to leverage MT output of suboptimal quality, for example through post-editing (PE)” (158). Similar to Jiménez-Crespo’s study, van Egdom and Pluymaekers focus on stakeholder groups with divergent experience levels and thus, given the fact that “the quality of a target text eludes the grasp of the receiver” (159), whether different stakeholders, including novice translators/students and professionals, perceive quality differently. And if so, considering they have diverse quality definitions, there seems to be a lack of consensus on what constitutes “optimal quality”. What are the common error types attributed to this kind of output?

For this reason, the authors open the discussion about quality by referring to David Garvin’s seminal text on Total Quality Management (1984), where he lists five definitions of quality he observed in different domains of knowledge, i.e., the transcendent, product-based, user-based, manufacturing-based, and value-based approach (159). The metrics and frameworks for evaluating translation discussed so far are part of the effort to avoid the transcendent, intuitive approach to translation. Since the 1990s, as a result of standardizing efforts for quality, analytical evaluation grids were developed and the processual requirements for quality services were set, e.g., ISO 17100: 2015; ISO 18587: 2017, as product- and manufacturing-based approaches entail. User and value-based approaches, on the other hand, served to answer the underlying needs of clients and end users (160). Yet, on this point the literature lacks empirical research into how end users perceive different degrees or levels of post-editing in a project-based setting, and the question the authors ask is very relevant: “Do they always prefer the most elaborate degree of post-editing, which is relatively costly, or

can lesser degrees also be fit for their purpose?” (160) Thus, in addition to different levels of post-editing, this study also tries to assess “sender image, attitude towards the product and purchase intention” (160). The participants of this study were selected either from potential customers of the sender of that text (e.g., the informative party) through convenience sampling of IT students studying at different universities in Denmark and graduates who worked in the software industry. Apparently, both groups had over-average domain knowledge. Before the test, which was done either with in pen-and-paper or online, participants received a short introduction about the topic and the target group of the text (161). The raw MT output they were presented was post-edited by fourth-year translation students. The participants were provided with instructions corresponding to the different degrees of PE including minimal, light, moderate and full PE (162). The instructions were based on levels or degrees of revision proposed by Mossop (2014) and the validity of these materials was ensured beforehand. Edit distance and the effort required for four different versions of PE was measured by a calculation.

Results demonstrate that in the case of the informative text, few manipulations were needed to get from light PE to moderate PE and from moderate to full PE, signaling a statistically significant difference. However, considerably greater effort was needed on the part of the post-editor, to leverage the quality from moderate PE to full PE (163). The difference between moderate PE and full PE, however, was never significant (166). In the case of the instructive text, PE from light to moderate did not require a significantly different effort for four independent variables, namely content, logic, usability, and purchase intention (166). But there was a significant difference between moderate and full PE for two variables: language use and style (167). Interestingly enough, the main question of this study – “*What is the effect of increasing degrees of post-editing on end users’ perceptions of texts, senders and products?*” (160) – receives a negative answer here as “full post-editing does not always elicit the most positive judgements from end users” (168). It is very clear from the assessment results the informative text elicited that, though the difference between moderate and full post-editing was never significant, the edit distance between the two degrees can be comparatively large (168). The authors did not differentiate age and gender among the

participants, but they suggest referring to other studies in which younger age groups displayed lower expectations in terms of text quality (e.g., Hedges 2009) (169). They further propose that measuring text quality perception of LSPs can be the first step taken for future research on PE quality and state that “text quality is context-dependent” and hence, “subject to a somewhat opaque process of constant (re)negotiation” as the additional effort required for higher degrees of PE may not result in more positive judgments about text quality (158).

Garcia’s study (2011) is a pioneering work, using “advanced leveraging”⁴⁷ methods provided by Google’s Translator Toolkit by default when it was used as a CAT tool and TM did not offer any fuzzy matches for certain segments. The author tests Allen’s (2003) hypothesis that post-editing would help translators produce significantly faster and higher-quality work compared to translation from scratch (218). In this second stage of experiments with the tool, the author adds two additional sets of tests with more participants. In the first experimental study, the results were contradictory to Allen’s hypothesis: Some students spent almost the same amount of time on full post-editing as on translating in a traditional way, while the quality of post-edited passages was no lower than those manually translated. Therefore, “[c]ontrary to Allen’s expectations, post-editing was not significantly faster and traditional translation from scratch was not better” (218). This study analyzes how SMT post-editing is influenced by language direction, text difficulty and the translator’s level of performance. Two cohorts of students were asked to translate from scratch and post-edit MT output in different orders. Participants were asked to translate to full quality, which was measured according to NAATI guidelines. After the projects were completed, the reviewers were asked to blind mark them. At the first attempt, translating by post-editing was faster in only 15/28 tests, equal to 54% of the cases and this small difference could not be considered statistically significant (220). At the second, participants were asked to translate from their ‘weaker language’ to their mother tongue and they were allowed to search for information

⁴⁷ By using the TAUS (2007) terminology “advanced leveraging”, Garcia here refers to the procedure where features developed in some TM tools are exploited to the most to facilitate reuse at the sub-segment level (2011, 230). For an in-depth review on the topic, Rothwell, Moorkens et al. 2023 and Forcada 2014.

on the web (for people translating from Chinese to English). This time, the difference in favour of post-editing could be considered statistically significant as post-editing MT was 62% faster (221). The author interprets this difference as “signalling that post-editing will increase productivity when working into the second language” and if the MT quality is high enough, the MT version may save translators time in the process of understanding the source text since it provides them with a draft (221). Therefore, the reasonable explanation behind these contrasting data can lead to a very general result that for some sentences “MT may have been more of a hindrance than a help” and for others it is useful enough to justify the money and effort invested in it. However, one issue might have clearly impacted the time spent on post-editing as in the second set of tests they had the chance to search online, which would obviously take more time (222). On the other hand, the difference in the perceived difficulty of the texts presented some very interesting results. When participants were asked to post-edit those passages which they perceived to be the most difficult, post-editing into the mother tongue was slower than translating from scratch but when they were translating into their weaker language, i.e., English, they were faster on average (222-23). In terms of differences among participants, there was no significant change in the way the best and the poorer people performed. Thus, the author argues that in terms of productivity, “(i) post-editing may serve strong performers as well as poor ones, (ii) it may be best suited towards less demanding source texts, and (iii) gains are likely to increase when translating into the second language” (224). In further studies, the author tries to ascertain if the results would apply again when translating into the mother tongue as well as into the second language, when dealing with a source text perceived to be more difficult and when the participant group changed (225).

With regard to first question about language direction, in 67% of the cases compared to 63%, where participants translated into their strongest language, they did better when post-editing. The author here notes that the difference would probably be higher if the participants had not been allowed web searches when translating from the source text, as they were not when they were translating into their stronger language (226). As for the difficulty of texts, the difference in post-editing and translation from scratch was highest when they were working on translating difficult texts into their weaker language and lowest when they were

asked to work on the most difficult passages, but in their stronger language (226). According to quality evaluations, both groups always performed better when post-editing but the results indicate that weak performers benefited more from post-editing compared to the stronger ones when translating into their stronger language, i.e., mother tongue “(in 26/32 passages, compared to 19/32 for the best performers)” (227). The difference in terms of quality was “more noticeable” in the case of the weaker group working into their mother tongue (227). Thus, the author argues that “the type of translator that Google seems to have in mind for its Translator Toolkit” is “the non-professional, or less skilled translator, working on simpler text, and into their second language” (228). Garcia proposes to repeat the same experiment with similar populations to test its validity and reliability if research on the same topic does not support the findings. Yet, his results are in line with O’Brien’s study (2007) where “[w]orking from MT matches has been found to require a similar amount of effort when dealing with fuzzy matches of around 80–90%”, and with Guerberof’s results, which indicate post-edited MT matches were found to contain fewer errors than the translated fuzzy ones (2009), and also with Fiederer and O’Brien’s study where post-edited segments were evaluated to be “as clearer and more accurate (if not as stylistically pleasant) as the conventionally translated ones” (228).

After Garcia’s trailblazing work, Marian Flanagan and Tina Paulsen Christensen’s study⁴⁸ on testing post-editing guidelines with translation trainees reflected on the divergent quality expectations and understandings of different groups of users and producers. They stress the fact that trainee translators enrolled in an MA course presented competency gaps which made it difficult for them to interpret PE guidelines for publishable quality, but this ultimate conclusion was also caused by the wording of the guidelines. Therefore, the authors propose training measures to address these competency gaps due to the fact that there is “a lack of PE training and of clear and consistent guidelines for post-editors” as well as translation trainees. (2014, 257). Referring to the results of a number of earlier studies which

⁴⁸ The authors highlight that their study is “the first to use baseline industry-focused PE guidelines in an academic setting”, a statement which reflects the gap between academia and industry in terms of best practices developed recently related to MT use (258).

investigated translation trainees' ability to identify issues in the raw MT that need essential changes, their speed and their ability to adhere to the PE guidelines, Flanagan et al. note that "more experienced translators are faster and more accurate post-editors than those with less experience". However, very experienced translators also had a tendency to make a higher number of preferential changes, while the less experienced translators seemed to make fewer changes in total (De Almeida and O'Brien 2010 quoted in Flanagan and Christensen 2014, 258). Another similar study by Depraetere (2010) aimed to identify PE guidelines that could be used in the classroom based on edits made or not made by her students during a PE task (259). The results show a major problem with students' analysis of errors, since they did not "rewrite correctly transferred sentences to make the text more fluent", did not edit terminology-related errors "even if a better solution was available" and did not notice information omitted in translation, and overall, most errors were not edited sufficiently by the students. Yet, the authors suggest that her training with "minimal instructions" might have influenced trainees "intuitive" PE choices (259). Participants of Flanagan et al.'s study were MA level translation trainees who had no MT or PE experience aged between 27 and 32. They were asked to machine translate and post-edit a five-page online journal text according to TAUS PE guidelines for publishable quality (2010) and after that write a six-page reflective report, followed by a second assignment which asked them to complete the same procedure for a marketing and information leaflet from a Danish pharmacy. The first project was aimed at medical experts, whereas the second one was aimed at laypersons (26-61). All trainees were supplied with a translation brief, which clearly presented relevant instructions, including information concerning function and the receiver of the target text and they were also interviewed three weeks after the second project via a semi-structured method (261). They were asked questions about how they interpreted each bullet-point instruction in the TAUS PE guideline for 'good enough' and 'publishable' quality, which the author analyzed in detail. The authors highlight that the main focus of their study was not measuring speed and translation quality but whether their research design was applicable and unbiased and also, whether they could point to problematic issues in PE guideline design (263).

In their analysis on the two guidelines, the authors classify each guideline as being either ‘problematic’ or ‘non-problematic’ for their own training context, “based on the trainees’ retrospective interview data and the supplementary sources”, and base their decision for ‘problematic’ on three criteria, which are:

(1) If at least two trainees fail to adhere to the guideline or ignore the guideline altogether. (2) If at least two trainees misinterpret the guideline, which might be due to trainee competency gaps, or the wording of the guideline. (3) If criterion 1 applies to at least one trainee and criterion 2 applies to at least one trainee” (263-64).

In the case of the introductory paragraph⁴⁹, the authors state that all three trainees expressed difficulties about how the stylistic elements mentioned were to be interpreted. For instance, one trainee described in her report how she spent the majority of her time making ‘unnecessary edits’ (or preferential changes). She also admitted to “perhaps not performing the PE task she was asked to do” because of these changes (264). Another one also expressed feeling compelled to make unnecessary changes “based on the purpose and function of the text, and the target audience”. In the interviews, they all said that they believed “aiming for a ‘stylistically fine’ TT conflicted with their translator training”. Furthermore, some suggested that they had difficulty “adhering to all the guidelines simultaneously, because they believed the stylistic elements of the IP conflicted with other guidelines” (264). Therefore, the authors considered this paragraph as problematic (264). In case of the second instruction which was interpreted by one participant to mean that post-edited text should be ‘grammatically correct to enhance the text’s credibility’, by one participant was comprehended similarly by the others; hence the authors considered it to be non-problematic (265). In the case of the third instruction on consistent terminology to be followed, two trainees did not refer to key terminology in their reports and when asked about it during the

⁴⁹ The TAUS guideline here states that for good enough quality the text must be comprehensible, accurate “but not being stylistically compelling”, while for publishable quality comprehensible, accurate and stylistically fine.

interview, they either expressed difficulties in interpreting this guideline or misinterpreted it or at least, could not “understand the importance of adhering to it”; hence the authors classified this instruction as problematic (266).

The fourth instruction on editing any offensive, inappropriate or culturally unacceptable content was also considered problematic by the authors as the participants suggested they had interpreted the instruction incorrectly, did not adhere to the translation brief or had preconceived ideas about the text type and therefore did not adhere to the instruction (268). The fifth instruction on using as much of the MT output as possible was also considered problematic as some participants clearly reflected on their unnecessary edits in their reports. Again, they could not fully understand what the concept of ‘style’ referred to here. All of them felt that this instruction was the most difficult for them. Based on the feedback, this instruction was also considered problematic (269).

The sixth instruction, stating that “basic rules regarding spelling, punctuation and hyphenation apply” was considered non-problematic, as almost all participants followed the standard usage (269). The final instruction on correcting formatting was considered problematic due to “the trainees’ lack of PE experience, and their unfamiliarity with translating different text formats” (270). The reports and the answers collected demonstrate contrasting results with Depraetere (2010), where the author asserted that there was “no need to (over-)emphasize the lack of need for stylistic changes or rephrasing”. On the same issue, O’Brien (2009) stated that in her study “the less experienced participants made fewer preferential changes than the more experienced participants” since they “felt that they needed to strictly follow the guidelines” (O’Brien 2009 cited in Flanagan et al. 2014, 271). The authors state that they lack the relevant data to compare their findings to both earlier studies. The authors propose that “renewed focus on the translation brief, and more emphasis on language skills” are definitely needed in new training programme and they also suggest designing “a set of PE guidelines for good enough quality, based on the approach taken here” and then, assessing the results through feedback (273). This study raised some very important questions about the translation brief including whether “all translator trainers define a translation brief similarly” and “all translator trainers emphasise the importance of the

translation brief” (271). The distinction between light and full PE, which surfaced in the descriptions “good enough” compared to “publishable” quality is analyzed in the empirical studies reviewed below.

Maarit Koponen’s article (2015) on her course design for MT and post-editing reflects on “skills that are specific to PE”⁵⁰ drawing on previous research by O’Brien (2002) and presents learning to evaluate MT suggestions including TM and the usability of segments which apparently refer to “learning to make only necessary changes and to discard suggestions that require too many changes” (4). The course objectives include theory of the history of MT and PE; practical use of MT and PE; controlled language and pre-editing for MT; post-editing guidelines and quality levels; quite interestingly, post-editing without source text and PE competences (4) are two separate topics covered in the course. The course had five assignments which focused on comparison of rule-based and statistical MT, pre-editing, post-editing without ST, quality levels and MT quality evaluation. At the end of the course, students were asked to write a reflective essay on their experience. Koponen states that she selected “relatively general interest texts that would not require particular expertise” to be edited and for the MT evaluation task, short and simple sentences for asking directions, making purchases etc. were taken from the tourist phrasebook dataset used by Rautio and Koponen (2013) and translated by three MT systems, the in-house systems developed by the European Commission, Sunda Systems and Google Translate (7-8). Koponen reports on some very positive results from the first assignment as students were able to analyze the types of errors found in each MT output and reflect them against their understanding of the type of MT used and they were also able to assess the use of MT and PE critically. Most students expressed doubts, particularly about the feasibility of PE without source text, based on their experience. Some said “MT provided a surprisingly good starting point” to work on and there was “not much” to edit and PE was not “particularly challenging”, while some felt “they

⁵⁰ This notion of PE requiring a skillset different from “human translation as well as revision of human translated texts both in terms of the cognitive processes and the practical goals and processes” is based on the earlier studies by Krings 2001 and O’Brien 2002 (Koponen 2015, 3). The existence of an international standard devoted to it also supports the argument. However, overused TM segments and revising and re-using them also causes a dilemma where the argument that PE is a unique practice seems no longer valid. Therefore, the empirical data on measuring effort is crucial more than ever.

probably would have translated the text at least as fast” (8). However, most students found it “somewhat difficult to differentiate between the PE quality levels, and to determine what changes were necessary for the ‘good enough’ quality” (8-9). Like Flanagan’s findings, after examining their own corrections, many considered that they were probably correcting “too much”. The exercise on examining their own PE times and edits was well-received inasmuch as comparing their PE times and corrections. They also commented that MT evaluation tasks concentrating on language only or meaning only issues “made them more aware of the issues” (9).

Students’ reflective essays were analyzed according to themes. 7 out of 13 students said they had very little knowledge about MT and PE before the class though they were familiar with at least one of the systems (9). 10 out of 13 “explicitly” stated that the quality of MT output was higher than they expected (10). Yet, 8 out of 13 said that they found it hard to trust MT, which was a point raised during PE without ST as students discussed whose responsibility it was to correct the final text (10). One common point that surfaced in the students’ essays is related to the theme of quality thresholds. For instance, some of the students identified issues “where the MT quality was likely too poor for any intervention to be fruitful⁵¹” (10). Based on their self-measurement of speed, students varied greatly and had different opinions on how they viewed their time issue (11). Some suggested that once they felt comfortable, they felt they were saving time compared to translation from scratch while most indicated they either would have translated the same text faster or that there was no difference in time, which is similar to Salmi and Koponen’s earlier findings (2014). Quite interestingly, a student wrote about the connection between experience and PE speed. Again, this remark might be interpreted as García’s results (2011) which demonstrated that “‘poor’ students benefited more from MT” maybe due to not having yet “formed effective translation skills” (11). Recalling Flanagan’s results some students felt they were on their “editing too much”, working on the translation “until they were happy with the final version”, an opinion which clearly had an impact on their understanding of speed. One student even went as far

⁵¹ Here the certain quality range refers to a level “deemed most usable for teaching purposes by Kliffer (2005): not bad enough to contain such frequent low-level errors that PE would be completely pointless, but not such high quality that there would be hardly anything to correct” (Koponen 2015, 10).

as to say “her final version should properly be called ‘inspired by the MT’ rather than ‘post-edited’”, whereas others tried to manage “only with minimal corrections” and did not attempt to “make the text their ‘own’”, yet they also expressed concerns about idiomatic MT output in their final translation (11). Very similar to what Salmi and Koponen reported (2014), the survey results indicate that there was an “apparent desire to edit the final translation to be as perfect as possible, even if it meant rewriting”⁵² in contrast to Depraetere’s research findings where students “‘did not feel the urge to rewrite it’, and did not rephrase the text if the meaning was clear” based on the analysis made by the author (2010).

As predicted from the above answers, 8 out of 13 students ‘explicitly’ stated that “they viewed MT as a tool, not a threat for them as translators” (12). Another common point in most essays was a clear understanding that “MT and PE are not feasible in every situation”, pointing to the argument that translators need to be critical users of technology since most students also argued that PE had some specific limitations particularly in the case of use without ST (12). Additionally, a number of them were optimistic about their career as future post-editors. As a final point, the author noted that in a future experiment she would reconsider the ordering of the assignments. Specifically, the valuation task on adequacy versus fluency might prove to be more fruitful, she argues, “in the beginning of the course, as it could assist the introduction of PE requirements and “good enough” quality, which the students found particularly hard to accept” (12). In that sense, comparing post-editing guidelines is the sensible approach to observing whether any specific guideline is better worded in the instructions that differentiate between light and full post-editing. For that reason, Ke Hu and Patrick Cadwell compare the five proposals⁵³, i.e., TAUS, O’Brien (2010), Flanagan and Christensen (2014), Mesa-Lao (2013), and finally, Densmer (2014), for post-editing guidelines published for wider use and not kept for internal use within corporate

⁵² This remark can be interpreted within the scope of unclear guidelines instructions as well as noted by Flanagan and Christensen (2014).

⁵³ The authors clarify that they selected general guidelines rather than those that are “language dependent or aiming at specific contents” (346).

bodies or international organizations. The authors argue that “[e]very post-editing project needs to have specific guidelines for translators to comply with”, otherwise clients and LSPs would not be able to set clear expectations (346). The table in which they summarized the

Figure 2: A Comparative Study of Light PE Guidelines presented by Hu and Cadwell (2016)

LIGHT POST-EDITING	TAUS (2016) (FIANAGAN & CHRISTENSEN, 2014)	O'BRIEN (2010)	MESA-LAO (2013)	DENSMER (2014)
Accuracy	TT communicates the same meaning as ST	Important	Important	Factually accurate
Terminology		No need to research	No need to spend too much time researching if incorrect	Be consistent
Grammar	May not be perfect	Not a big concern	No need to correct unless the information has not been fully delivered	Correct only the most obvious errors
Semantics	Correct			Correct
Spelling	Apply basic rules	Apply basic rules		
Syntax	Might be unusual	Can be ignored	Do not change	
Style	No need		No need	
Restructure	No need if the sentence is correct		No need if can be understood	Rewrite confusing sentences
Culture	Edit if necessary	Edit if necessary		
Information	Fully delivered			
Others	Use as much raw MT output as possible	Textual standards are not important; very high throughput expectation; low quality expectations	No need to change a word if correct	Fix machine-induced mistakes; delete unnecessary or extra machine-generated translation alternatives

key points of the descriptions provided by these guidelines can be found below:

In Densmer’s proposal, the authors note, the key phrases are “factual correctness” and “good enough”. Light post-editing also requires turning a blind eye to ‘minor’ errors. Yet, in full PE, because absolute accuracy is aimed for, the effort to achieve human-level quality in PE “may exceed the effort to have it translated by a linguist in the first place” (Densmer 2014 quoted in Hu and Cadwell 2016, 348). Here the dilemma in the industry is revealed as she exposes the fact that “many clients want the quality of full post-editing with the price and speed of light post-editing” (ibid.). The points summarized in the table also reflects the

contradictory expectations for accuracy, terminology and culture. For instance, all the proposals consider accuracy important, but they also see no need to conduct research on terms and neither do they find it necessary to guarantee that information translated is complete and fully delivered apart from the first two proposals. Densmer insists that terms used, though not checked in terms of accuracy, need to be consistent. For TAUS, Flanagan and Christensen (2014) and Densmer (2014) the quality of content being semantically correct is also important. Furthermore, O'Brien and Mesa-Lao believe that there is no need to spend too much time researching and correcting the incorrect terminology. But three guidelines argues that basic spelling rules should be applied and text should be edited according to target culture's needs if necessary, or even adapted. Yet, if the post-editor observes a very clear grammar mistake this should not be considered a big concern and does not necessarily need to be corrected. O'Brien stresses that the quality expectation from light PE is low and textual standards are not deemed important.

Figure 3: A Comparative Study of Full PE Guidelines presented by Hu and Cadwell (2016)

FULL POST-EDITING	TAUS (2016)	O'BRIEN (2010)	FLANAGAN & CHRISTENSEN (2014)	MESA-LAO (2013)	DENSMER (2014)
Accuracy	TT communicates same meaning as ST	Important	Important		Absolutely accurate
Terminology	Key terminology is correct	Key terminology is correct	Key terminology is correct	Apply the term as used in the term database for any incorrect terminology	Consistent and appropriate
Grammar	Correct	Accurate	Correct	Correct	Correct
Semantics	Correct		Correct	Correct	Correct
Punctuation	Correct	Apply basic rules	Apply basic rules		Correct
Spelling	Apply basic rules	Apply basic rules	Apply basic rules		Correct
Syntax	Normal		Correct		Make modifications in accordance with practices for the TL
Style	Fine	Ignore stylistic and textual problems		Not important	Consistent, appropriate and fluent
Restructure			No need if the language is appropriate	No need if the sentence is semantically correct	Rewrite confusing sentences
Culture	Edit if necessary	Edit if necessary	Edit if necessary		Adapt all cultural references
Information	Fully delivered	Fully delivered	Fully delivered		
Formatting	Correct	All tags are present and in the correct positions	Ensure the same ST tags are present and in the correct positions;		Correct (including tagging)
Others	Basic rules apply to hyphenation; human translation quality	Apply basic rules to hyphenation; high throughput expectation; medium quality expectations	Use as much raw MT output as possible; ensure the untranslated terms belong to the client's list of 'Do not translate' terms	No need to change a word if it is correct; accept the repetitive MT output	Perfect faithfulness to the source text; fix machine-induced mistakes; delete unnecessary or extra machine-generated translation alternatives; cross-reference translations against other resources; human translation quality

On the other hand, the expectations from full PE are quite to the contrary as the table above summarizing Hu and Cadwell's key points demonstrates (2016). There are many overlaps including issues related to terminology, grammar, semantics, punctuation, spelling, syntax, culture and information. But some very clear differences are also observed. For instance, TAUS and Densmer clearly expect human translation quality. They also stress "fine", "consistent, appropriate and fluent" style. Yet, O'Brien and Mesa-Lao do not consider

style to be important. O'Brien even suggests ignoring stylistic and textual problems. These two guidelines expect medium quality after full PE. Mesa-Lao argues that as long as it is correct there is no need to change a word. Densmer, on the contrary, advocates for cross-referencing translations against other resources. In that respect, there is a very transparent division of thought in terms of quality between scholars and the industry⁵⁴, the latter being "more inclined to human translation quality" (351). Therefore, the question follows: if the quality expected from full PE process is similar to translation from scratch, will post-editing bring any productivity gains? Research⁵⁵ indicates the answer is not necessarily yes. But the present study reviews those studies in the next section, and continues with linguists' interpretation of guidelines and the behavior observed in relation to guidelines, specifically over-editing.

Jean Nitzke and Anne-Kathrin Gros's study (2021) on over-editing behavior reveals results that would also pave way for research on post-editing. The authors start by referring to an earlier study by Mellinger and Shreve (2016), who analyzed experiment data collected from nine translators' attitudes to revising TM segments where participants changed 60% of exact matches which did not require any charges but changed only 74% of the fuzzy matches, even though all of them needed change when given complete freedom even to delete an entire segment. Many of these unnecessary changes were considered preferential changes on a syntactic or lexical level. The explanation given for this behavior by the authors was "because the participants do not have to create their own translations when there are fuzzy and exact matches but instead have to create compare source and target exact, they actively look for mistakes" (21). There is also the possibility that the participants might have their own translations in mind, resulting in two competing versions of the target text segment (21-22).

⁵⁴ Lee Densmer was a senior manager at Moravia, one of the largest LSPs globally which was acquired by another LSP, i.e., RWS, when she published her PE guideline in her blog at the Moravia website (Hu and Cadwell 2016, 348).

⁵⁵ On that point, Hu and Cadwell refer to studies by Guerberof (2009) and Federico et al. (2012) who reported productivity gains in their research, whereas Gaspari et al. (2014) indicated that post-editing could lead to productivity losses compared to translation from scratch" (351).

The same behavior was also observed in relation to PE in De Almeida's study on the classification of errors in the post-editing of an IT text where the main categories were essential changes, preferential changes, unimplemented changes and introduction of errors (2013). 'Preferential changes' here were considered synonymous with over-editing. In a text with 1008 words, participants made 45.16 preferential changes on average, which means an unnecessary change was made in every 22.32 words (22). These two studies signal the same issue that was observed in the comparison of PE guidelines. As the participant is asked to keep the editing process to a minimum, how they interpret this instruction or how they decide on a threshold between minimal and over-editing is unknown. Moreover, given that these translators are not necessarily trained in PE, they may not have any experience/practical knowledge on how to lower their quality standards and correct only the essential mistakes, as well as disregarding their personal style and habits (22).

On that note, Nitzke and Gros analyzes over-editing or preferential changes in PE and revision scenarios focusing on datasets from three studies. In the first, participants were asked to translate from scratch, light post-edit and full post-edit either technical or medical texts. The output was provided by statistical Google Translate and 150-words long. In each task a different text was used. The participants were advanced translation students and the instructions they were given were adapted from the TAUS PE guidelines (23-24). In the second study, the ST included four newspaper articles and two sociology papers. Half of the participants were professional translators and half of them were translation students with little professional experience. They were asked to translate two texts from scratch, bilingually post-edit two MT outputs and monolingually post-edit two MT outputs, meaning without the source text at hand. The length of the texts varied between 100 and 150 words and the MT system used was again Google Translate. Participants were asked to edit the texts minimally and retain as much raw MT as possible (24-25). In the third study, the translations created from scratch in the second study were adapted for a revision task. Higher quality translations were selected⁵⁶ and each text contained 100-150 words. These texts were manually manipulated by inserting natural errors found in human translations in the previous sessions

⁵⁶ The selection criterion here was being "flawless regarding errors and style" (25).

of the experiment. A total of three to five errors was inserted into each text and no sentence contained more than one error (25). Each text contained errors belonging to different categories to prevent participants from recognizing a certain pattern. The error types inserted include orthography, grammar, sense, omissions, consistency, and coherence. The authors of the study state that “[e]very error type occurred with equal frequency in the texts” (25). Participants were asked to revise six translations with access to ST. The brief below was also provided:

A colleague has translated the following texts from English into German for a renowned German newspaper (print) or for an encyclopaedia has asked you to now revise the translation before it is published. Please concentrate on correct content and language and please insert your directions directly into the texts. You will get paid for seven minutes per text. There is no internet connection—please revise the text without Internet research. (25-26)

The authors report that the full PE exercise in the first study took the most time, whereas the participants in the second study took the least time. This result was something expected since the quality demanded from the full PE was the highest, which probably also caused more characters to be inserted into the text rather than erased (26-27). For the first study, the authors categorized over-editing instances and quantified them. Nine areas defined for over-editing consisted of lexicon, syntax, style, addition, deletion, grammar, spelling, punctuation and insecurity, which denotes deleting a target text unit and inserting the same unit without any changes (28). For the last two steps, the authors identified instances of over-editing, under-editing, and necessary editing and again categorized the over-editing instances. In the first phase of analysis, when the changes were counted, the difference between light and full PE was “less striking than expected” (28). Yet, the difference in total changes was still statistically significant (29). Moreover, the differences between necessary changes and over-editing instances were on average much clearer in full PE compared to light PE with about 70% of the changes characterized as necessary and about 30% as over-editing. In light PE,

the difference between necessary changes and over-editing instances is not statistically significant with distribution being half to half (29).

In the second study, the difference between necessary and over-editing instances were significant and when the results are compared to the full PE task in Study 1, the authors observed the participants made “significantly more necessary changes than changes that could be characterized as over-editing” (29). In the third study, quite interestingly, almost the same number of inserted mistakes were corrected and also not corrected. Therefore, the difference between corrected and uncorrected inserted mistakes is not statistically different. Moreover, an average of 4.39 translation items that were not manipulated were altered in every session. As a result, only about 31% the changes were necessary, whereas the remaining 69% were unnecessary which can apparently be categorized as over-editing (29). Quite similar to these results, in Mellinger and Shreve’s study (2016), 60% of the exact matches were edited, though they would not necessarily have need editing, while 26% of the fuzzy matches which would have required changes in the translation were not edited (29). Most of the changes characterized as over-editing were made in the revision study, i.e., the third study, followed by the light PE if we consider the distribution of changes between necessary ones and over-editing. However, in the case of the total number of changes, most changes were introduced in full PE with light PE and revision following in order. The results can be interpreted with regard to the experimental situation as “the participants might have felt the need to edit more instances per text than were actually to deliver good results” and may have expected to have many more errors included in the texts (29-30).

The authors note that over-editing instances were based on lexical or stylistic preferences. The authors interpret over-editing in the light PE task to be caused the vagueness in the PE instructions. For the revision study, 8.7% of over-edits were categorized as preferential syntactic changes and 4.5% were unnecessary changes according to raters, indicating the fact that preferential changes occurred more frequently in the revision task (30). The frequency of additions and deletions was almost the same except for the revision task. When the results for light and full PE were compared, additions and deletions occurred twice as much in the full PE task. Yet, the total numbers of additions and deletions were quite

low compared to other over-editing instances (31). Nitzke and Gros conclude the study by stressing that “[t]he final text simply needs to be acceptable under the quality requirements, not perfect” in PE and thus trainees need to be taught about suppressing personal preferences in order to make both PE and revision worthwhile and lucrative (32-33). After all, in PE in particular, efficiency might be more important than the overall quality of the final text (33) because apparently “translators lose most when they put more effort into the task than they are paid for” (32). Furthermore, they point to future research on the influence of experience in revision and PE on over-editing behavior since obviously assumptions about whether translators with greater translation experience over-edit more often or over-edit less often needs to be tested, as both cases seem plausible (33). Also, they suggest analyzing whether the under-editing⁵⁷ instances caused by more fluent and natural-sounding MT output would also make sense (Nitzke and Gros 2021,33). At this point, we turn the discussion about difficulty of evaluation where, as White et al. (1994) attest, “evaluation must exploit intuitive judgments while constraining subjectivity in ways that minimise idiosyncratic sources of variance in the measurement”. Here evaluator training comes to the forefront for the consistency of TQA to avoid cases of final decisions “boiling down to the opinion of one person over that of another” (DePalma and Kelly 2009).

Before addressing the issue of evaluator training for more accurate evaluations and the accompanying issue of evaluator agreement, an issue about automatic evaluation, which is widely used for its plus sides in terms of budget and efficiency, needs to be addressed. Though the main purpose of this study is mainly comparing statistical, phrase-based and neural system, Burchardt et al. (2017) reveal the fundamental key issues about automatic error annotation. The authors argue that before 2017⁵⁸, because statistical MT systems are not assessed using broadly-defined test suites, their performance depends greatly on the particular input data, parameter settings, and etc. which makes it difficult to draw final conclusions about them. Also, because there is no strictly “correct” MT output as there is no

⁵⁷ Nitzke and Gros suggest that under-editing also needs to be analyzed by assessing and categorizing data on (2021, 33).

⁵⁸ The year this research was published is quite important as it dates back to the widespread use of NMT and LLM systems.

clear equation such as $X+Y=Z$ in translation evaluation, MT as a type of machine learning task cannot be compared to those tasks of NLP such as parsing or fact extraction “where the expected results can be more or less clearly defined” (160). But still, MT output is automatically evaluated by comparing the output to reference translations or post-edited MT output, e.g., Hjerson. Therefore, research mostly focuses on maximizing scores on test corpora that are comparable to the training/tuning corpora, which would yield unreliable results in other settings and domains (160). For all these reasons, the authors suggest testing system performance on error classes.

Using the test suite⁵⁹ they built for QT21 project, which has segments selected from various parallel corpora and other sources including grammatical resources, e.g., the TSNLP Grammar Test Suite (Lehmann et al., 1996), and lists of typical translation errors found online, they compare different versions of Google Translate, different commercial and non-commercial neural systems (162) and they evaluate only the phenomenon in the respective sentence and disregard any other errors that could be annotated, which gives them the opportunity to determine whether a translation error is linked to the phenomenon under examination or not. In the former case, the segment is deemed ‘incorrect’ and the error is not considered to be related to the phenomenon, i.e. the error issue, and is counted as correct (162). The linguistic categories used in this study are composition, function words, long distance dependency (LDD) & interrogative, multi-word expressions (MWE), subordination and verb valency, phrasal verbs and idioms. As the study is conducted on the German-English language direction, the results are limited; nevertheless, the RBMT system performs the best of all systems on average, achieving the highest number of correct translations (83%) on verb tense/aspect/mood and verb valency, while one NMT system demonstrated quite contrasting results for the different categories, compared to RBMT (167). The latter system was quite successful on the remaining categories including coordination & ellipsis, LDD &

⁵⁹ The authors state that the latest version of the test suite comprises a collection of about 5,000 segments per language direction which are classified in about 15 categories (most of them in both directions) and about 120 phenomena, most of which are similar with some language-specific ones as well (Burchardt, et al. 161).

interrogative, MWE, NE & terminology, special verb types and subordination (167). Thus, as a final argument, the authors propose using hybrid combinations or pipelines for future research where the ultimate goal is to automate test suite testing by manually providing positive and negative tokens for each test (169). The usability of such a method for automatic post-editing remains untested as far as we know.

A similar study presenting data on whether automatic metric and human evaluation results align in a comparative setting where phrase-based statistical (PBSMT) and neural systems are assessed was conducted by Castilho et al. (2017). The domain of texts is educational as the study was done as part of a European-funded project focused on the translation of MOOCs, and four language pairs, i.e., English into German, Portuguese, Russian and Greek, were tested (117). Referring to earlier work⁶⁰ on comparing two similar systems, the authors highlight that they “performed side-by-side ranking, adequacy and fluency rating, post-editing and error annotation based on a predefined taxonomy” with professional translators (118). The training data were obtained from various sources including WMT shared translation tasks, OPUS, TED from WIT3, a corpus of Coursera MOOCs, and authors’ own educational data (118). In the ranking task, 3 experienced professional translators were given a subset of 100 source segments with their translations from PBSMT and NMT in a random order and they were asked to rank the segments using Google forms in the first task and post-edit the outputs in terms of adequacy and fluency using a Likert scale of one to four and do error annotation in the second (119-20). The automatic evaluation metrics used were BLEU, chrF3 and METEOR (Banerjee and Lavie, 2005) and two⁶¹ post-edited texts were used as references for each segment. Additionally,

⁶⁰ Referring to Bentivogli et al.’ study (2016) in which HTER was used to “estimate the fewest possible edits from pre- to post-edit”, Castilho (2017a) state that the results indicated technical post-editing effort (in terms of the number of edits) decreased by 26% on average when using NMT compared to the best-performing PBSMT system. There were significantly less word order errors, specifically in case of verb placement and fewer lexical and morphological errors in the NMT output (117). Another study by Wu et al. (2016) suggests that NMT strongly outperforms the other approaches and only “some additional ‘tweaks’ would be required before NMT would be ready for real data” (Castilho 2017a, 117).

⁶¹ Popovic’ et al. (2016) indicate that using a single post-edited reference from the evaluated MT system tends to introduce bias (120).

the authors note “HTER metric (Snover et al., 2006) was used to estimate the fewest possible edits between pre- and post-edited segments” (120).

For the rating task, the evaluators were asked “How much of the meaning expressed in the source fragment appears in the translation fragment?”– with one in the case of adequacy meaning ‘none of it’ and four was ‘all of it’, and in the case of fluency, one meaning ‘no fluency’ and four meaning ‘native’. The assumption the authors had about NMT was high ratings for fluency and possibly lower ratings for adequacy specifically for longer segments (see Cho et al., 2014; Neubig et al., 2015). In the second task, participants were asked to post-edit the MT segments to publishable quality and then error annotate the texts based on a simple error taxonomy comprising inflectional morphology, word order, omission, addition, and mistranslation. The assumption here was that there would be fewer morphology and word order errors with NMT, especially for short segments (120). The results of automatic evaluation differ as BLEU, METEOR and chrF3 scores considerably increase with NMT compared to PBSMT for some language pairs. Also, regarding the amount of PE effort, the HTER scores indicate that more PE was performed when using the output from the PBSMT system for some language pairs. No statistically significant differences for HTER scores were found, while chrF3 scores showed some good improvement for NMT over PBSMT for some languages but no statistically significant improvements for others (121). When it comes to human evaluation, NMT was rated as more fluent than PBSMT for all language pairs and the NMT systems appear to have fewer problems when compared against the PBSMT systems for all the languages (121). In terms of adequacy, on the other hand, the results were not completely consistent (122). Overall, the NMT output was mostly found to be more fluent and comprehensible, though not without errors (124). In error annotation, the total number of issues was greater for PBSMT than NMT for all language pairs. Furthermore, the number of segments left without error annotations meaning no errors were found is greater for NMT across all language pairs, as NMT output was found to contain fewer word order errors and fewer inflectional morphology errors in all the target languages (124-25). The authors stress that, interestingly, the percentage of errors found in both systems seems to follow a certain pattern where

inflectional morphology, word order, and mistranslation were the most frequent. Therefore, the numbers indicate that the output from both systems are “to some extent dependent on the particular language pairs involved, and are clearly influenced by the specific morphosyntactic features of the target language”, which causes the PE effort to “inevitably” vary from one target language to another while keeping the other two parameters, i.e., text type and the domain constant (125). Interestingly enough, no clear improvement regarding omission and mistranslation errors was observed between PBSMT to NMT.

In the case of ranking, there was a clear preference for NMT across all language pairs and for different content types (65% for the business analysis content forum, 54% for the medical training transcript, 52% for the physics transcript and 55% for the advertising transcript). The inter-annotator agreement values show moderate agreement (125). The authors argue that in settings where fluency is considered to be more important, NMT seems to score better, therefore NMT may be “a good fit” for the subtitling domain in general, specifically for content that is not particularly specialized (126). Thus, in terms of PE effort, fewer NMT segments were considered by participants to require editing in the MT post-editing task (127). Analyzing the keystroke measures and the minimum number of edits required to go from pre- to post-edited text, the results for technical effort demonstrate a certain decrease for NMT in all language pairs. Yet, in terms of the cognitive effort, though the results are not statistically significant, NMT segments required more effort compared to PBSMT segments. Participants’ feedback also indicates a similar result with some expressing that they found NMT errors more difficult to identify. In contrast, “word order errors and disfluencies requiring revision were detected faster in PBSMT output” (127). Yet, productivity⁶² does not seem to be positively influenced by the use NMT as none of the participants reached the average rate of professional throughput, i.e., 0.39 words per second in the language pairs studied. But the authors speculate that this result might be caused by participants’ unfamiliarity with the interface, the specialized nature of the texts and the

⁶² A common generalization among translators is that they achieve productivity with continuous work and often report that their productivity peaks half-way into their day (Castilho et al. 2017a, 128).

annotation tasks they were required to do after the post-editing task (127-28). Thus, the authors argue NMT “may not as yet offer more than an incremental improvement in temporal PE effort” for production (128). These results shed light on a major discussion about preferential changes and whether all the annotators agree on the error issue type that is at hand for a specific problem. The next section analyzes the studies that focus on both these dilemmas.

3.4 Preferential changes and inter-annotator agreement

Preferential changes caused by over or under-editing has a long history behind it considering that both the instances are caused by inherent quality evaluations. How translators, translator trainers and regular people with varying experience levels in translation tend to categorize error issues and their severity under different categories has been discussed so far based the results of empirical studies. This part expounds on the phenomenon of preferential changes and their causes, which contribute to low inter-annotator agreement (IAA).

Starting with Lommel, et al.’s preliminary endeavors in the QTLaunchPad project (2014) using “detailed, in-line human-generated quality annotations” where agreement is low, the authors suggest that the problems observed in annotation are “likely to be endemic (although covert) in quality evaluation for both machine and human translation”. On examining the factors that have an impact identification and classification of errors, they find the ones below “salient”:

(1) disagreement as to the precise spans that contain an error; (2) errors whose categorization is unclear or ambiguous (i.e., ones where more than one issue type may apply), including those that can be described at different levels in the taxonomy of error classes used; (3) differences of opinion about whether something is or is not an error or how severe it is” (1).

In that case, a specific example would suffice. In WMT shared tasks in 2013, IAA between two or more annotators did not exceed 0.40 (κ) and *intra*-annotator agreement, which refers to consistency of raters’ own ranking patterns when doing the same assessment task multiple

times, surprisingly did not exceed 0.65 (Bojar et al., 2013, 6–8). In order to be considered a reliable measure, it is required to have an agreement of at least 0.85. From a functionalist perspective, translation quality is always relative to given specifications or the given job. That's why TAUS MQM enables custom issue types to be introduced if necessary. Additionally, factors such as resource availability, production environment, target audience, etc. are effective in determining whether a certain translation is considered correct or not. For instance, when an on-demand instant MT system is used, quality can be determined by the output's success in permitting the reader/user to accomplish a task. In that case, low levels of Accuracy and Fluency can be forgiven for the usability of output.

Human input is required in one of four ways inside the loop: for human-generated reference translations, MT output rating based on perceived quality, MT output post-editing, i.e., implicit error markup and error annotation, also called “explicit error markup” in some sources. Based on the assumption that the human input explained above provides a reliable basis, automatic evaluation tools use these sources of input to measure the quality. Lommel, et al.'s study has an experimental setup where professional translators from commercial LSPs evaluated a set of 150 sentences using WMT 2012 shared task (2014, 2). Before selecting the sentences for annotation, human evaluators reviewed the MT output for the 500 translations of each of the systems – SMT, RbMT and a hybrid MT – and the 500 reference human translations, according to a scale of 3 ranks where 1 means “Perfect output (no edits needed)”, 2 “Near misses (1–3 edits needed to be acceptable)” and 3 “Bad (more than 3 edits needed)” (2). From the Rank 2 sentences, the authors pseudo-randomly selected 150 sentences, to create a “calibration set”. After uploading this corpus to the translate5 system, the annotators were provided with a set of written guidelines and invited to a webinar to introduce them to the tool and the task (2). The main error issue types were Accuracy covering Terminology, Mistranslation, Omission, Addition and Untranslated and Fluency covering Register/Style, Spelling, Typography, Grammar and Unintelligible. The results of this study are enlightening since they also give spans for errors after examining the positional tagging in pairwise comparisons (3). For instance, in the case where Rater 1 and 2 are quite similar ($\kappa = 0.89$) and Rater 2 differs from them, both Rater 2 and 3 identified the same errors and they

disagreed on the precise span, causing lower κ scores. Moreover, for the language pair EN>DE the average kappa was 0.30, which means the results can be considered fair (2014, 3). Overall, the IAA scores for the human annotation task were lower than those for the rating task, with the highest scores in the annotation task ironically enough corresponding to the lowest scores in the rating task (4). On that note, the authors speculate on the reasons for these “discouraging” results and hidden problems in many translation quality assessment tasks/methods (4). The first problem they address is related to the precise scope of error, i.e., the span. In some cases, although the raters agreed on the issue type, they disagreed on the span that the errors covered. The instructions clearly asked them to annotate the minimal span⁶³ referring to “*only* the issue in question” but some reviewers annotated the entire verbal unit instead of the verb alone (4). But the fact that at the sentence level, they reached at least partial agreement in terms of error issues which makes assessing IAA hard. Furthermore, word order had the highest percentage of instance in the error annotation, which can be interpreted as the confusion about the location of an error (4).

The second problem with annotations is related to possible issue types. Since the exact nature of the problem is not clear from the text, one annotator marks a sentence as having a specific issue, while another annotator does not mark the same issue type. Lommel et al. disregarded the location of the error within the segment for his study and considered annotators to be in agreement if two annotators annotated the same issue class as occurring in the segment. Overall, in terms of specific issue types *Mistranslation* and *Terminology*⁶⁴ revealed high levels of confusion among the annotators. The *Function word* category also demonstrated very high confusion, as can be expected with very different profiles between the language pairs. The total numbers show that it is one of the most frequently occurring

⁶³ A study which asks annotators to do manual evaluation on MT output for specifically focusing on the error span states that even in a study where IAA is high—fine-grained error classification was not one of the objectives—many evaluators were not sure about how to point to the exact span of errors. Thus, deciding on the error span was considered to “very subjective, and therefore cannot be precisely guided by instructions” (Popovic 2020, 5066).

⁶⁴ In a similar study on error annotation Burchardt et al. 2014 discusses this specific confusion and argue that *Mistranslation* was used for longer error spans in general, whereas *Terminology* was used “primarily for single-word spans” (Lommel et al. 2014, 5)

and problematic in the entire corpus. On the other hand, *Word order* revealed high levels of agreement between annotators, even if the span-level agreement was significantly lower. Ironically enough, the instances where some annotators did not spot an error while some said there was an error had a relatively high percentage (2014, 5). Based on these results, the authors point to confusion within the hierarchy, because annotators were instructed that if no issue applied precisely at one level, they should select the next highest level. Therefore, many issues could be annotated at multiple levels in the hierarchy particularly in the cases where the precise nature of the error was not entirely crystal-clear (5-6). The authors also state that due to the guideline writers' assumption of some linguistic knowledge from the annotators, a simple list of issue types and definitions along with some general guidelines were "insufficient" as most annotators knew how to correct an error "intuitively" yet did not exactly know how to classify a certain error in an analytic framework. Here the authors advise the guideline writers to formalize and proceduralize the decision-making process (6). The fact that annotators might have different opinions⁶⁵ on what constitutes an error, a fact that causes biases in research⁶⁶, has a significant impact on IAA. Setting aside the convenient conclusion about different opinions being welcome, Maja Popovic' proposes that to overcome the long hiatus of automatic and semi-automatic evaluations, language-related issues must re-emerge in annotation task. The major points that need to be addressed in future research in that respect are definition of error issue categories beyond traditional linguistic ones, generalization of error issue classes, correspondences between two types of error categories and lastly, automation of issue identification, e.g., pre-annotation (2018, 152-53).

In order to compensate for the lack of linguistic knowledge of annotators the previous study results yielded, Marta R. Costa-jussà and Mireia Farrús propose using precise guidelines formulated with linguistic principles in mind where each instruction "strictly mark"s the difference between the subfields of linguistics including morphology, syntax,

⁶⁵ Here the discrepancy between automatic and human evaluation must be addressed. The authors state that for the assumption that "good enough" evaluation scores can be achieved with automatic evaluation is only valid for limited cases where the training data used for MT must be similar to a great extent to the reference translations. Yet, with heterogeneous training data and references, biases are unavoidable (Lommel et al. 2014, 6).

⁶⁶ For a similar study on a case disagreement do not constitute "noise" or "error", see Popovic 2021.

semantics, and orthography (2015, 157). Interestingly enough, the study results show high inter-annotation agreement thanks to wide-error coverage. The authors hypothesize that the main objectives of an ‘implicitly’ linguistics-based approach to evaluation would be high inter-annotation agreement, a linguistic classification of the translation output errors, and evaluation without a reference apart from the usual new information to benchmark the MT systems (158). The major drawbacks of such an evaluation, on the other hand, would be the requirement for bilingual annotators and the time-consuming nature of the process (158). Because the authors were aiming for higher IAA using non-expert annotations⁶⁷, which is the main motivation of these guidelines (158), they trained four MT systems, Apertium, Translendum, Google Translate and UPC, with the medical corpus provided by the Universal Doctor project, which aims at facilitating communication between healthcare providers and patients (162). The guidelines prepared had four main branches: orthographic (related to language writing standardization); morphological (dealing with internal structures of words and their modifications); semantic (the meaning of individual words and combinations, and how they combine to make a meaningful sentence); and syntactic (whether words are combined in a grammatically correct way). Annotators were asked to classify any error they could find into one of these categories (158).

The subcategories of the major error categories include “no correspondence between source and target words, non-translated but necessary source words, missing target words, and non-translated proper nouns or translated when not necessary”, as well as polysemy, homonym, and expressions used in a different way in the two languages under semantic errors (160). Syntactic errors include all the errors found related to prepositions, relative clauses, verbal periphrasis, clitics, missing or spare articles in front of proper nouns, and syntactic element reordering (160-61). Under morphological errors, annotators were asked to list instances of lack of gender and number concordance, apocopes, verbal morphology (inflection) and lexical morphology (derivation and compounding) related errors, and morphosyntactic changes (159). Finally, orthographic errors deal with punctuation marks,

⁶⁷ The authors here refer to previous research where non-expert annotations “if produced in a sufficient quantity, can rival and even surpass the quality of expert annotations, often at a much lower cost (Su et al., 2007; Snow et al., 2008)” (Costa-jussà and Farrús 2015, 158).

erroneous accents, letter capitalization, joined words, spare blanks coming from a wrong detokenization, apostrophes, conjunctions and errors in foreign words (159). Three non-expert annotators were asked to rate the sentences using a Likert scale from 1 (good) to 5 (bad) in terms of adequacy and fluency (162). The weighted kappa for IAA was found to be 0.75, which is good according to (Landis and Koch, 1977). The reason for such a high kappa value is attributed to the accurate design of the linguistic guidelines (163). Correlating the results for adequacy and fluency at the sentence level, all the equations followed a statistically significant trend. For this annotation framework, adequacy was designed to clearly correlate to semantics, “a little to syntax and nothing to orthography and morphology because these two levels do not interfere in the ‘understanding’ of the translation”, while fluency correlated with all levels in the order of importance, with the major ones being semantics, syntax, orthography, and morphology (164). The total errors for adequacy were 0.58 and for fluency 0.57 in the final correlation, and thus the level of agreement achieved was found to be surprisingly high (164).

Finally, in one of the earliest studies on IAA⁶⁸, Stymne and Ahrenberg attempt to increase the agreement by implementing a number of changes to their original research design (2012). With these changes, they refer to variables that might impact the correlation of levels of inter-annotator reliability and the properties of the error taxonomy, measures to improve inter-annotator reliability, comparing the performance of the two systems and, lastly, using a corpus of short sentences instead of a random sample of sentences” (1785). For the experiment setup, the authors asked two authors to perform an error analysis of two statistical MT systems trained on the Europarl corpus (Koehn, 2005). The annotators were very familiar with the error annotation tool, i.e. BLAST (1786). The system is modularized, enabling error typologies to be customized and the error span of each annotation could be stored in memory. The annotators were given only a few instructions to follow. For instance, one was asking them to make as few changes as possible if there was more than one

⁶⁸ Bayerl and Paul’s study on IAA reveals that agreement varied based on domain, the number of categories in the annotation scheme, the training received by the annotators, and calculation method (2011).

possibility and if a word was annotated with several errors including both in meaning and form, all of them required to be registered and also, some specific instructions such as one about prepositions where in the case of a preposition was translated by a wrong preposition, the error should be classified as a Disambiguation error (Word sense) if the meaning changed; otherwise it would be classified as Wrong function word (1786).

Stymne and Ahrenberg designed the experiment in two phases, with 50 sentences in each phase. In the first phase, the annotators discussed the error taxonomy given to them which had no example-based guidelines. After the first phase, the annotators wrote down guidelines based on their discussion. In the second phase, they used these guidelines to annotate two sets of sentences, one short, where sentences contained a maximum length of 20 words and a random set where the sentences had 58 words maximum (1786). All in all, the guidelines had twenty-three instructions under nine different general categories. The results reveal that the agreement was higher in phase 2 than in phase 1 and the discrepancy between the phases was larger when more complex annotations were used rather than simpler annotation schemes (1787). Based on the data collected in phase 2, the authors analyzed which types of distinctions between error types that were easy or hard to make by using errors the annotators both agreed and disagreed on (1787). For most errors the annotators either agreed on the class (266 instances) or only one annotator identified the error (287 instances), indicating that the instances where the two annotators marked the same error but chose a different error category was quite rare (1787). In the case of mis-matching errors, the differences were mostly within the top categories. The most confused error types were Wrong syntax and sense. The results did not correspond to the ones presented by Popovic' and Burchardt (2011) where the choice between incorrect lexical choice and missing or extra words was "especially difficult" as there was no confusion between Wrong sense and Missing or Extra words" (1787-88). The only instance where a specific mismatch stood out was for the linguistic categories, in which the distinction between syntax and sense was most common, corresponding to 48.5% of the instances" (1788). However, the results Stymne and Ahrenberg present confirm the findings of Bayerl and Paul (2011) who suggest inter-annotator agreement is higher with roughly grained having fewer categories and when

annotators receive more training (1789). Furthermore, the reason for differences in error categorization, the authors indicate, are annotators' varying views on how an error should minimally be corrected. In that light, developing guidelines asking for minimal corrections, and explicitly storing the intended target hypothesis with a specific instruction would increase IAA. Yet, using corpus of short sentences for annotation did not yield higher IAA, thus refuting the initial hypothesis in that respect (1789).

This review covered experimental setup with numerous variables. Yet, the major variable that promises higher quality MT output is apparently domain, a topic which is going to be covered in the next part.

3.5 Domain and PE effort: Evaluation of NAATI specialized texts and “non-specialised domains”

Besides evaluation and quality estimation, three parameters that are of concern in the case of MT output quality are language combination resource, the size of linguistic corpora for training and also, the domain. Domain is a generic name and commonly misused. For this reason, the scope of the term must be more specifically delineated. For instance, ISO 17100 defines ‘domain’ as “subject field, sphere of knowledge or activity having its own specialized culture, social context, and linguistic characteristics” (2015, 3). *The Cambridge Dictionary of Linguistics*, on the other hand, offers a more abstract, philosophical definition under the entry ‘domain’ but here quoting only from the 4th and 5th definitions is explanatory enough:

4. The social context (including, for instance, participants, location and topic) that influences choice of language, dialect and style. 5. In Conceptual Metaphor Theory, a relatively stable packet of information, a ‘knowledge structure’, about some area of human experience that is required for the interpretation of a given lexical word. A distinction is drawn between basic domains deriving from the physical world as experienced and conceived by humans and abstract domains that are more distant from the physical world, such as psychological states, human relationships and intellectual concerns such as number, truth and falsity, scientific theories (2013, 141).

The ISO definition of the term in the translation services context is derived from these definitions but presents a more concrete, practical understanding. Domain names encountered in the websites of multilingual language vendors such as automotive, life sciences, legal etc. after all showcase certain packets of information and accompanying knowledge structures, hence a certain register. These specific linguistic properties are assembled under the heading ‘text-type convention’, which ISO17100 defines as “subset of specifications for the target language content related to the content type and domain” (2015, 3). A note to this entry is illuminating as it states “[f]or example, with acts of law, there are formal conventions always used by legislators” (3). The commonly accepted categories of domain entail linguistic elements attributed to register and style. Merging the definitions provided with ISO17100, register is better explained as “variety of language used for a particular purpose or in a particular social or industrial domain” (ISO17100: 2015, 3). Apparently, both terms act like different sides of the same coin. Further to that, the answer to the question whether a certain MT system has the computational capacity – capacity here refers to MT architecture and training corpora – to relay these elements in another language lies in corpora tagged with certain domain and sub-sentential parsing (see Lavie et al. 2008). In line with this question, ISO 18587 mentions domain as a decisive element in the decision to use MT output in a translation project “whether the source language content is suitable for MT and subsequent post-editing or not since combined MT/post-editing efficiency depends on the MT system, language combination, domain and style of the source language content” (ISO 18587: 2017, 5).

Apart from being one of the primary factors in the decision to use MT output, the standard on post-editing articulates on “domain competence” as part of post-editing competence, defining it as “the ability to understand and master the domain content produced in the source language and to render it in the target language” (7), identical to the domain competence as one of the professional competences of a translator that 17100 reports on (2015, 6). Additionally, the term “domain terminology” is used four more times (6, 8, 12 and 14) throughout the document. Another is in Annex D titled “Client–TSP agreements and project specifications” where as Part of Project specifications, the document lists “linguistic

specifications for the product, consisting of standard and optional specifications” and give some examples of standard linguistic specifications such as “compliance with client and/or domain terminology, lexical cohesion and phraseology, etc.” (12). The final one is in Annex E on automatic post-editing where as part of objectives of the process the standard has to “adapt the output of a general-purpose system to the lexicon/style requested in a specific application domain” (14). Overall, domain and the language register specific to it are pivotal aspects for a higher quality MT output and how the industry approaches the issue in terms of each localization project is very clear. The rules for project managers’ tagging of domains were set a long time ago. But because of language combination resource limitations including data crawling and pre-editing rules, domain is not considered a sub-parameter that needs to be researched.

Apart from the industry’s static approach to domain, the National Accreditation Authority for Translators and Interpreters (NAATI) offers a certification system accepted in Australia. They offer two certifications for translators, i.e., one for recognized practicing translators and another for certified translators (“NAATI Certification System”). The prior one is provided if there are no tests available in a language pair. For interpreters, though, there are several certificates, including ones for recognized practicing interpreters, certified provisional interpreters, and certified interpreters with a higher-level generalist interpreting test. Besides the certified conference interpreter, one can apply for two additional specialist tests which are certified specialist health interpreter and certified specialist legal interpreter tests. Apparently, the categorization used in NAATI certifications distinguishes between two domains possibly due to the fact that translators/interpreters working in other domains such as automotive, marketing, IT etc. do not necessarily need a specialist certificate to prove their experience and adequacy to get jobs. In the next part, the testing results for comparing sentence and document-level evaluation will be discussed.

3.6 Sentence-level evaluation or document-level evaluation?

One significant criticism made for MT evaluation is whether evaluating output segment by segment results in misevaluation, if the evaluators were given the whole document alongside

the segments, the evaluation would be less biased and inter-annotator agreement scores would be higher. Attempting to answer these questions, Sheila Castilho’s study on comparing “the inter-annotator agreement (IAA) scores, the effort to assess the quality in different document-level methodologies, and the issue of misevaluation when sentences are evaluated out of context” (2021, 34). Yet, here lies a critical issue: when attempting to situate segments which are generally either a complete sentence or a shorter piece of text in a larger content, it is still unclear how much of the text needs to be shown for understanding the suprasentential context (34). In other words, the definition of document-level is ambiguous. Not much was known before Castilho’s study about using document-level methodologies to assess MT quality even though some previous works focused on the issue (Läubli et al., 2018; Toral et al., 2018; Barrault et al., 2019⁶⁹; Castilho et al., 2020⁷⁰). By ‘different methodologies’ Castilho means comparing inter-annotator agreement (IAA) scores in three settings, which are random single sentences, sentences in context, and full document scores (34). Her results demonstrate significant difference between two methodologies. When the evaluators assessed individual sentences within the context of a document in comparison to a methodology with random sentence assessments, a good level of IAA was achieved, whereas when the evaluators gave each document a single score, IAA scores were low. Moreover, they also noted that there was a repetition of misevaluation cases in the random single sentences evaluation scenario (34). But before moving to Castilho’s findings, we need to see the results of previous studies.

In Scarton et al.’s earlier study (2015), where they analyzed six of the most commonly-used automatic machine translation evaluation metrics, e.g., BLEU, TER, METEOR, to provide labels for quality estimation at document and paragraph levels, results showed more post-editing was performed in paragraphs, which suggests “several issues could

⁶⁹ In *Proceedings of the Conference for WMT19*, a number of studies compared document-level and sentence-level models and for one such studies’ results, researchers express that “document-level performs significantly worse than the sentence-level model (Talman et al. 421). The next year in *Findings of the WMT20*, Barrault et al. explains that one major category among several dimensions that the test sets differed was either to use document-split or sentence-split data (2020).

⁷⁰ Castilho et al. aims to identify the necessary context span in a study involving three domains and 18 target languages (2020). Castilho and Knowles (2024) offer the most comprehensive survey of context in NMT and its evaluation to our knowledge.

only be solved with paragraph-wide context”, revealing “important discourse dependencies in translation that go beyond QE” (127). In Toral et al.’s study (2018), evaluators were better able to evaluate the translations when given access to more context and also IAA between professional translators was higher compared to that between non- experts and the authors also suggest that “[h]uman evaluations should proceed taking the whole document into account rather than evaluating sentences in isolation” (121). Moreover, in Läubli et al.’s study (2018) in the case of fluency, document-level set-up had higher IAA than sentence set-up but for adequacy the opposite was the case (4795).

Returning to Castilho’s study in terms of adequacy, random sentence (RS) score has higher IAA than both the document-level scenarios, i.e., where translators had access to full document and were asked to either give one score per individual sentence, hence sentence-in-context score (SC) or one general score for the whole document, hence document score (D) (2021, 34). For the fluency assessment too, RSs scenario has higher IAA than both-document level scenarios, SCs and Ds. But the SCs scenario shows slightly higher IAA for DeepL output (36). Additionally, the Ds scenario, in general, demonstrated lower IAA scores than RSs and SCs methodologies, while the RSs scenario shows slight higher IAA than SCs (38). Overall, the highest scores were given to a document-level methodology where translators can tag errors for each sentence with access to the full document, which can result in better IAA (38-39). Moreover, – interestingly enough– the IAA scores are generally lower when full texts were ranked (Ds), compared to IAA scores given where translators rank individual sentences with access to full texts (SCs) (39). When it comes to the post-task questionnaire done for assessing the effort⁷¹ spent on evaluation, all the translators gave positive answers for the SC scenario which means they found it easier to understand both source and translation when evaluating sentences in context, found it easier to recognise

⁷¹ In a post-task questionnaire evaluators/translators were asked rate the effort they spent on RSs and SCs scenarios. Effort was assessed on a Likert scale from 1 to 6, where 1 is a negative answer, meaning very difficult up to 7 with 8 meaning very tiring and 9-10 meaning strongly disagree, and finally, 6 is an affirmative answer, meaning very easy/not tiring at all/strongly agree (Castilho 2021, 36).

adequacy and fluency problems, as well as spotting errors and choosing the best translation when having access to full texts. Finally, they also expressed being more confident with their assessment (36). The author writes they also observed misevaluation issues in this study, where disagreements in the RS scenario are more often related to ambiguity and lack of context (41). On a final note, Castilho speculates about why Rs assessment has a higher IAA, saying “the reason methodologies with random single sentences show higher IAA agreement is because raters tend to accept the translation when adequacy is ambiguous but the translation is correct, especially if it is fluent” (42). This might explain why the scores for adequacy in the document-level scenario are divergent, while the scores for fluency are rather homogeneous. After all, NMT systems might have probably improved fluency, resulting in misevaluation being more likely to happen in an RSs set-up. Referring to earlier work on document-level evaluation and context span (2020), Castilho states the results of this recent study corroborate the previous results where “IAA scores for document-level reaches negative levels, and the level of satisfaction of translators with that methodology is also very low” (42). Yet, again how much context is necessary for a random single sentence evaluation is still not clear, as can be observed from the confusing results.

In short, this confusion might stem from both error annotation frameworks⁷² and the method of assessing inter-annotator agreement. On that note, Lommel states that for building a more time-saving, accurate, rational, non-biased and hence, lucrative method for annotating errors, in building MQM they chose to have a “more parsimonious” and “task-appropriate” approach instead of “very fine-grained distinctions” (2018, 115). In the case of assessing inter-annotator agreement, Amidei et al.’s study titled “Is agreement definitely overrated?” reveals that “most data collected for NLG evaluation fail the reliability test” (344). The authors explain this lack of reliability with tests due to “irreducible human language variability” and suggest that using correlation coefficients and agreement coefficients together “obtain a better assessment of the evaluation data reliability” (344). With this final

⁷² Vardaro et al. even suggest that frameworks for error annotation mislead taxonomic research (2019).

criticism on the reliability of inter-annotator agreement, we now turn to studies focusing on the question of whether post-editing is more efficient than translation from scratch.

3.7 Quality over quantity?

The fundamental reason why the need for the technology known as machine translation to be developed is the expectation that a computer would translate much faster, making the translation/localization process cheaper and the content would be accessible in many more languages and, most importantly, the output would have human parity quality. The industry's adaptation of MT has been quite fast in the past few years. But on the contrary, as more research is conducted and published it is revealed that post-editing might take more effort and time than translation from scratch. In one of the recent articles published on the topic, Silvia Terribile states that her results “indicate that (a) PE is usually but not always faster than HT; (b) average speed values may be misleading; (c) translation speed is highly variable; and (d) edit distance cannot be used as a proxy for post-editing productivity, because it does not correlate strongly with speed” (2024). She also stresses the fact that most previous studies on post- editing (PE) speed focus on “evaluating machine translation (MT) systems’ performance, rather than the efficiency of PE itself”. In terms of average translation speed values, studies reported large variability and revision speed was almost always overlooked. For instance, some LSPs use edit distance to measure the productivity gains obtained through post-editing and/or to determine the amount paid to linguists (ELIA et al. 2022) but a number of studies have reported conflicting results about the correlation between technical and temporal PE effort (see Cumbreño and Aranberri 2021; Moorkens et al. 2015).

Ironically enough, as more LSPs invest in MT and “post-editing has become the dominant translation method” according to Memsource (2020), the process of full post-editing has come to a position where it is an alternative to human translation. Pushing forward with it, ISO 18587 does not mention revision as a key step for any type of PE and also the literature reviewed in the previous chapters assumes that revision is only the case in a workflow with human translation and the PE process does not involve a revision step. In one of the very rare studies, Özlem Temizöz reports on comparing the two workflows where

subject-matter experts post-edit MT output and professional translators subsequently revise the translation in the first, and vice versa in the second one, professional translators post-edit the MT output and subject-matter experts do the revision. The means of comparison in this study, as with the rest of the studies on the topic, are temporal, referring to the time needed for post-editing, e.g. speed expressed as a words per hour ratio, and technical effort, referring to the all the operations needed to practically carry out PE, namely “deletion, insertion, reordering, or a combination of these operations” (Krings 2001, 179) and measured through edit distance . She states the experiment with engineers and translators who passed the error threshold yielded the following results:

On average, translators spent 62.70 minutes for post-editing and 27.40 minutes for revision [...] On the other hand, engineers spent 55.66 minutes for post-editing and 33.31 minutes for revision. Unlike the mean duration of post-editing, in which engineers worked faster than translators (55.66 minutes against 62.70 minutes), the mean duration of the engineers’ revision is longer than that of the translators’ (33.31 minutes against 27.40 minutes). In revision, engineers processed, on average, 18.6 w/pm, while translators processed 29.2 w/pm (Temizöz 2017, 8).

But when the outliers were excluded and mean values were taken, there was no significant difference between translators and engineers in terms of speed in revising one another’s post-editing. In other words, neither group was significantly faster than the other, be it engineer post-editing revised by translators or translator post-editing revised by engineers (18). In terms of quality though, they observed a difference. Temizöz states that “engineer revision significantly increased the quality of translator post-editing”, while translator revision also increased the quality of engineer post-editing, yet the increase here was not significant. The final result is even more revealing. When they penalized recurrent errors, “the quality improvement brought about by engineer-revision of translator-post-editing is significantly higher than the quality improvement brought about by translator-revision of engineer-post-editing” (18). When recurrent errors were not penalized, neither parties’ quality improvement significantly differ (19).

Returning to the point about pricing in terms of temporal and technical effort⁷³, the question of whether they are interrelated has not been studied enough according to Terribile (2023). She quotes the results of “only four small-scale studies” she knows about. For instance, Sharon O’Brien found that speed and edit distance “correlate well” in the post-editing of SMT (2011, 19), whereas Moorkens et al. (2015) could identify strong correlations between them. Macken, Prou and Tezcan (2020) observed weak to moderate associations between speed and edit distance in a similar setting and a moderate association in the post-editing of NMT output with Cumbreño and Aranberri (2021) reporting that they observed only weak correlations in the post-editing of NMT output. Yet, no previous research has investigated this correlation of efforts at the revision stage of post-editing or human translation. Therefore, she analyzes data provided from an LSP, TranslateMedia, on human translation and full post-editing with almost the same group⁷⁴ of linguists in a language combination. The author investigated eleven language pairs with the highest number of translated words, starting from 2019 as the specific LSP adopted this technology in 2018. The translated content belonged to the genres of marketing (63% of content was human translation, 86% was PE) or technical marketing, referring to “marketing texts with a high percentage of technical terminology” (Terribile 2023). One essential drawback of the data was caused by the CAT tool used standing inactive over 2 minutes, resulting in the deletion of all time spent inactive, guaranteeing that breaks were not included. Thus, time spent researching and reading any document besides the source documents over two minutes was not considered. There is also a chance that some values could be incorrect, mostly due to technical issues or to linguists working offline. Edit distance was measured using the Levenshtein algorithm through the CAT tool with the formula “Edit distance = minimum

⁷³ It is not feasible for LSPs to take cognitive effort into consideration as there is no method to measure it without costs due to using eye-tracker tools and software.

⁷⁴ 92% of the post-editors also work on HT. However, as the amount of texts translated through HT is much larger compared to post-edited texts, only 22% of linguists working on HT do post-editing as well” (Terribile 2023, 4-5).

number of edited characters / number of reference characters” and it is expressed as a percentage. In other words, the lower the percentage, the fewer the edits.

Terribile focuses on three types of edit distance, i.e., post-editing distance (PED), PED including TM matches and repetitions (‘PED-TMR’) and revision distance (2023). This classification aims to target an accurate evaluation of revision speed as post-editors are assigned to edit a combination of MT output and TM matches, with some of the fuzzy matches and their edited versions are auto-propagated. This method will also let the researcher measure the distance between the translated and revised texts in addition to investigating whether a correlation between translation speed and edit distance can be observed and also, if edit distance can, in fact, be used as “a proxy for productivity”. In terms of results, only translation speed outliers were disregarded in the final calculation since all edit distance outliers appear to be valid. If a linguist works offline, this affects translation speed value, but not edit distance. The results demonstrate that PE does not always necessarily increase speed in all language combinations even in the cases where a significant amount of post-editing work has been completed in a language pair. Moreover, average revision values were higher in PE than in human translation in 8 out of the 11 language pairs, with an overall speed increase of +38%. This striking difference between human translation and PE at the revision stage can be explained as a result of the same quality standards applied to the revision of PE and human translation and also, the assumption that “the revision process is not supposed to present significant differences whether carried out as part of a PE or HT workflow” (2023).

On this note, Terribile hypothesizes that this might be due to the perception among post-editors that post-editing is inferior to human translation, hence they can be “less accurate” and complete PE revision jobs more quickly and the second reason could be “the quality of PE outputs might be frequently higher than that of human translation outputs, enabling linguists to work faster on PE revision” (2023). This second reason can be interpreted as a result of various factors. For instance, because the source texts processed through PE at this LSP are “typically ‘well-structured [and] often repetitive’ texts” (Kosmaczewska and Train 2019, 169 cited in Terribile 2023), those segments with PE might

be “more straightforward to translate” compared to revising human translation. But there is puzzling evidence about speed: When the mean values included outliers “on average, linguists have saved 18% of their time through post-editing”, yet when outliers were excluded “on average, linguists⁷⁵ took 4% longer when post-editing”. Even in the case where all outliers were considered valid, in only 11% of the jobs PE would have been very time-efficient. In contrast, PE has been counterproductive on average in the remaining 89% of tasks where “the average PE speed is 520 WPH, which is lower than the average speed achieved through HT, both including outliers (622 WPH), and excluding them (544 WPH)”. As a final note, since great variability was observed in the number of words processed in one hour with a high difference across language pairs, the standard deviation (SD) values without outliers also provided a similar results; as Terribile states, “SD is much higher in revision compared to translation, it is overall 87% higher in PE than HT [human translation] at the translation stage, and 35% higher in PE compared to HT in revision”. Given this high variability Terribile argues that it would be “beneficial” for LSPs – it would also make more sense in terms of pricing, i.e, the amount paid to freelancer linguists – “to consider variability also in the amount of time spent to process a text of a given number of words”, not vice versa. Therefore, she analyzes the number of hours spent to process a fixed number of words, which she set to 1,000, and the results are statistically significant: “H/KW dispersion is overall considerably lower in revision than in translation, and much lower in PE compared to HT”.

Overall, the correlations identified in Terribile’s research are partially in line with previous research on the interrelation between speed and edit distance in NMT post-editing. For instance, among the interrelations identified previously are “moderate correlations between speed and edit distance comprising changes to the MT output and TM matches (Macken, Prou and Tezcan 2020)”, and “weak correlations when considering changes to the MT output only (Cumbreño and Aranberri 2021)” (2023). But their results presented limited findings because they did not measure the speed and edit distance interrelation both including

⁷⁵ The highest speed of reading for a post-editor in this range was theoretically up to 19,800 WPH. The author suggests if this value is halved for reading both the source and target texts (although their lengths may differ), post-editors could read up to 9,900 words in an hour” (Terribile 2023).

and excluding TM matches. Also, parallel to the results of previous research (e.g. Moorkens et al. 2015; Parra Escartín and Arcedillo 2015; Toral, Wieling and Way 2018), high variability among the average speed obtained by different linguists was observed, which can be explained as a result of linguists' skills and practices. Also, the dispersion for average WPH values by different linguists was much higher in PE than HT whereas the opposite was the case for human translation/PE jobs. Finally, Terribile reports on "a moderate tendency" for PE as post-editing speed decreased when linguists made more edits to the MT output and a "weak tendency" when corrections to TM matches and repetitions were included in the edit distance values. The results are also in line with "a lack of enthusiasm from translator users" reported by Plitt and Masselot 2010; Guerberof 2012 even though MT is considered a beneficial translation aid by LSPs.

This chapter reviewed and discussed the major topics related to post-editing from both researchers' and industry's perspectives. Starting with the definition and different types of post-editing, empirical research into light and full post-editing was analyzed in detail, as well as the question of editing threshold. Next, it delved into the problem of the most accurate and reliable method of evaluation. Finally, it focused on studies on the correlation of speed and edit distance. The next chapter investigates the priming effect in the context of MT evaluation and post-editing.

CHAPTER 4 THE PRIMING EFFECT AND EXPERTISE

Empirical research on post-editing, inter-annotator agreement and error annotations reveal one thing clearly. Inter- and intra-annotator agreement is linked to several factors such as experience in translation and in a specific domain, as well as access to term searches, clear instruction on the type of post-editing expected from the translators and furthermore, what it entails in terms of instructions/guidelines. Kalle Konttinen et al.'s probing into the expertise level fitting for post-editing or revision is enlightening as it unmasks an assumption not only in industry circles but also among researchers and trainers. They reveal that post-editing is “sometimes envisioned as a task for beginning translators or even non-translators familiar with the subject matter” while revision is ‘often’ considered to be “a demanding task more suited to later stages of translator education and for more experienced translators”, referring to previous studies by Way 2009; Mossop 2020 (2021, 194-95). However, in a study by Guerbero Arenas (2014) results did not demonstrate any significant differences in the post-editing competences of experienced translators and translation students and in some other studies by Zaidan and Callison-Burch 2011 and Mitchell 2015, non-translators were even observed to have potential as post-editors in specific situations (Konttinen et al. 2021, 195). Yet, Yamada Masaru's (2014) research with foreign language students, who were specifically chosen as the opposite category to translation students) could not provide overall post-editing quality comparable to professional translators, as O'Brien argued that post-editing is a specialised task for translators (O'Brien 2002). This assumption about expertise in research can be interpreted as corresponding to the classification between “fit-for-purpose” or “good enough” quality, where emphasis on the correctness of information content can be prioritized over fluent language, which therefore requires no revision at all.

In quite a smart diversion from the established line of research, Daems and Macken's study where participants were asked to post-edit human translation and revise MT output the hypothesis that there would be fewer errors found during the revision phase because participants would be less critical and more over-revisions in the revision phase was not confirmed (2021, 68). In fact, there were more changes taking place with the MT output in

the revision phase and there were numerous instances of preferential changes but over half of the errors were uncorrected in both situations (68-69). Thus, the question remains how we can explain the editing behavior on scale between under and over- or hyper-editing. In this respect, this chapter analyzes what is called the priming effect in the post-editing process. When a participant is exposed to certain stimuli repeatedly, they tend to develop a familiarity with and therefore, not respond to it as they did in the first place. In our context, it causes errors to go unnoticed during post-editing. In the first part, the details of a well-known earlier psycholinguistic experiment by J. Kathryn Bock will be analyzed, followed by a brief review of priming in the second language context and finally, how it was unmasked in translation and editing research will be dealt with.

4.1 What is the “priming effect”?

In linguistics as well as psychology, priming can be briefly defined as “the tendency for speakers to reuse recently experienced structures” (Dell and Jacobs 2016, 212), and there is apparently more than one type, including structural, also known as structural repetition, and syntactic priming. The latter is explained as the cognitive process⁷⁶ in sentence production causing a speaker to be more likely to use a specific structure rather than a reasonable alternative when the speaker produces or comprehends a sentence with a particular syntactic structure (Ferreira 2009, 701). As more recent research indicates structural priming is not just a temporary change to the system, but rather a form of implicit learning (Dell and Jacobs 2016, 212) and therefore, the question underlying this study about MT output functioning as a type of foreign language ‘norm’ in a workflow where it gets post-edited by an inexperienced translator who can easily prime the translation norms is worth consideration from a linguistic approach.

⁷⁶ The cognitive process here comprises “the activation or strengthening of information” which implies a quasi-neurological energizing, excitation, or threshold reduction that persists over time, increasing the probability that the activated or strengthened information will influence subsequent cognitive processes (Anderson, 1983; Posner, 1978)” (356).

One of the pioneering studies on priming was done by J. Kathryn Bock (1986), who conducted three experiments with university students. In the first one, participants were shown target pictures which were explained with either a transitive or dative sentence in active and passive voice read aloud (360) and the participant repeated it aloud immediately after the experimenter. Next, they were shown filler pictures described with filler sentences. After this study phase, they were asked to identify the pictures and sentences they had seen in the study list by saying “yes” or “no” (363). After being given feedback about the accuracy of the answers, participants turned the page with the target pictures and described what was happening in the picture and made another yes or no decision. Following this recognition test, participants were asked questions about what the purpose⁷⁷ of the experiment could be, stressing their awareness of relationships between the pictures and the sentences and the effects that the sentences they repeated might have had on their descriptions of the pictures. They were questioned about special characteristics of the sentences that they repeated, and deviations in their picture descriptions from their normal speech patterns (363). The results of the first experiment suggest “[s]yntactic repetition appears to occur even when substantial differences in word order and grammatical roles distinguish the forms used to express semantically comparable messages” (366). Furthermore, students’ answers to post-experimental questions revealed that they did not think the sentences were in any way related to the pictures and they were also unaware of the similarities between the forms of the priming sentences and picture descriptions.

In the second experiment, the same procedure was used with a couple of differences. For the purposes of learning more about the penetration of conceptual representation into syntactic mechanisms, the research variable focused on the human and non-human agents in the passive sentence formulations. Then, the same set of transitive priming and target sentences were replaced by human and nonhuman agent sentences. Some drawings were changed to have more diversity among the human agents and patients. This time, though, all

⁷⁷ In order to minimize subjects’ attention to their speech and its structural features, the priming trials were embedded in a recognition memory test that puzzle participants’ understanding of the purpose of the experiment,” (359-60).

of the filler and target pictures were presented on slides and the filler and priming sentences from the study and test lists were recorded on cassette tape by a female speaker. The procedure was the same apart from minor changes. For instance, the subjects controlled the presentation of the slides and if a subject repeated a sentence from the test list incorrectly, the experimenter read the entire sentence again from a prepared script and asked the subject to repeat it once more (370). The most significant result of the second experiment was that “the repetition effect for passives persisted across variations in the messages conveyed by the priming sentences” (373).

In the final experiment, the procedure used in the experiment 2 was employed with only one change. Instead of showing slides one after the other, four slides were reversed to make half of the agents of the human agent events appeared on the left, and half on the right, and the same was done for the nonhuman agent events (374). The results of the final experiment were similar to those of the second experiment with one exception. That is, the increase in the passive utterances after passive primes occurred not only for nonhuman agent events, but also for human agent events (375). Similar to previous post-experiment questionnaires, none of the participants were “aware of any regular syntactic deviations from their usual speech patterns” (376). The results of the experiments were consistent with the hypothesis that states that “activated information corresponds to procedures that create grammatical structures in sentences” and can at least to an extent be separated from the conceptual features (383). These results are in line with studies on post-editing and revision effort where “the persistence of communicative intentions or discourse strategies that trigger the same syntactic forms, or to communicative tactics that predispose the repetition of structure, such as maintaining a question in active memory while formulating an answer” (359), hence causing the PE effort to decrease as the priming takes place. In the next section, the literature on whether these translation norms are primed through second language learning will be reviewed.

4.2 Priming effect and L2 learning

The priming effect has been tested in the context of bilingual reading research⁷⁸ for a long time, where researchers experiment with various parameters from the demographic characteristics of participants, such as age of acquisition and gender to different aspects of bilingual lexical representation and lexical decision-making including human and non-human agents, actives and passives, cognates⁷⁹ and non-cognates, word pairs, a faux amis etc. One such study that focuses on “masked translation priming” explain the process as where “the presentation of prime stimuli is masked so that participants are aware only of the language of the target items” (Davis et al. 2009, 137). In other words, the general procedure for priming which requires words from one of the participant’s languages (primes) to be presented prior to words in the other (targets) that require a response is altered and the prior stimuli are not explicitly marked or stated by the researcher. Generally speaking, though, the data used in across-language priming procedures involve sets of primes and targets that are, in fact, translation equivalents. The crucial aspect in the case of the priming process regarding bilingual lexical representation is assuring that the pattern of same-script masked translation priming effects⁸⁰ is “well established” (139). How it connects to post-editing research obviously because the process of constantly matching source and target entities priming inevitably occurs as the CAT tools permit the setting where “the processes set off by the prime can rapidly influence those used in processing and generating a task-relevant decision to the target” (Davis et al. 2009, 151) Yet, are these sets of equivalents used in priming studies limited to concrete lexical choices or are they also concerned with more abstract coupling?

In a recent study which investigates the degree of semantic alignment between translation equivalents, Chaouch Orozco et al. (2023) studied data obtained from tests with 200 highly proficient Spanish–English bilinguals in a translation priming paradigm consisting of 314 translation equivalents that varied in concreteness (2). Questioning the

⁷⁸ Experimental studies on conceptual, semantic and syntactic priming processes can be found in Trofimovich and McDonough (2011).

⁷⁹ Cognates refer to words which have a direct common ancestor as opposed to non-cognates which do not share a common ancestor.

⁸⁰ Davis et al. notes that “masked priming provides a window into the very earlier stages of word processing (it is presumed to tap pre-existing memory representation” (2009, 151) and thus, age of acquisition is the significant criterion in selecting participants for their study.

bilinguals' representation of meanings for words, the authors evaluated the extent of semantically mediated translation priming in concrete and abstract translation pairs (ibid.). Referring to earlier work where the results demonstrated inconsistencies in translation equivalents not entirely matching their meanings, the authors state that there is now a separate line of approach where equivalence is considered to be "a collective agreement by which interlingual communication is made possible, rather than an accurate reflection of how these words' semantic information is represented in the minds of bilingual speakers" (2). Thus, the authors suggest non-cognate translation equivalents⁸¹ as a good testing ground for the study of bilingual semantic processing in which the hypothesis "with higher overlap, more of the activation reaching the prime can spread over to the target, priming it more" could be tested (3). In other words, priming is assumed to be larger in a setting where semantic activation is more robust. The procedure of this study consisted of trials where a fixation point (500 milliseconds) was set, followed by the prime (200 milliseconds), and the participant received the target immediately after that and made the lexical decision with the target still on the screen. Attention checks were implemented pseudo-randomly (7). The participants were asked to respond as fast and as accurately as possible (8). The authors conclude from the data obtained the asymmetry of the emergent concreteness effect since "the less detailed semantic specifications of the L2 concrete primes could not activate the semantically richer L1 targets to the same extent as the L1 primes did with the L2 targets" (10). Similar to these findings, the comparable size of priming for abstract translation pairs in both directions is interpreted as "similar integration/representation of L1 and L2 abstract meanings in the bilingual lexicon" (10). Following this line of thought, studies on priming level and MT output post-editing will be reviewed in the next part.

4.3 Priming effect and MT

⁸¹ The authors briefly define these sets of equivalents as "translation pairs with no overlap at the form level" (Chaouch Orozco et al. 2023, 3).

Syntactic and lexical priming could be observed in the experiments conducted before. However, there is no clear evidence in the conclusions of the aforementioned studies on post-editing and theoretical studies on post-editing related skills of which specific skills are (could be) primed under the varying groups of subcompetences. For instance, do novice translators use MT output instead of looking up online dictionary entries and searching for terms, for learning and gisting about the content context or as half-done translations to be edited. Furthermore, since terminology is the category in which most errors are observed, would providing a custom built-in glossary be effective for the best post-editing performance or would it be more productive to expect freelance translators to search for terms themselves and evaluate their research output?

In line with these questions, Zou, Carl et al.'s study (2022) on the impact of translation brief and search conditions based on translators' experience is revealed results that could be interpreted in relation to the priming effect. The authors' initial statement that since a translation brief "specifies the intended audience and purpose of the translation in the target language", it is "meant to bias (or prime) translators, to activate particular, but not other, 'bodies of thought' (Gutt, 2004, 13) that answer to specific translation expectations" (62). Additionally, PE guidelines correspond to translation briefs, another prime for translators working on PE, specifying what needs to be post-edited for a particular project. In earlier similar research, including Daems and Macken's study (2021), which compared the quality of the translations after revisers' and post-editors' editing and Onishi and Yamada's study (2020), which compared the search behaviors of professional and novice translators, the results indicate that the revisers outperformed the post-editors when they edited MT output and the post-editors outperformed the revisers when they edited human translations in the first study, and "professional translators devote a higher percentage of time and operations into searches" as there was a high correlation between accuracy errors and the frequency and depth of searches in the latter one (Zou, Carl et al. 2022, 62). Also, in Daems and Macken's study, accuracy errors remain underedited in every case scenario, suggesting that different conditions influence the performance of the revisers and post-editors.

Returning to Zou, Carl et al.'s experimental setup, two groups of translators, whose L1 was either simplified Chinese or Japanese and whose L2 was in both cases English in both groups, post-edited four texts used for English-to-Chinese and English-to-Japanese ATA exams on welfare, war, tourism and racism and each text being around 250 words long – around and 10 segments each (2022, 62-63). Errors were divided into six types: “Mistranslation”, “Usage”, “Terminology”, “Grammar”, “Omission/Addition” and “Other” and the former four types were further evaluated as “Critical” and “Minor” depending on their severity (65). The authors grouped “Mistranslation”, “Terminology”, and “Omission/Addition” errors under the label “Accuracy”, and “Usage”, “Grammar”, and “Other” errors were grouped under the label “Fluency” for an easier comparison of variables (65). The evaluation was done on the Trados-Translog interface for aligning the keystroke and gaze data. The results demonstrate that the most frequent errors for Google NMT and post-editing annotation done by two professional translators were, respectively, Fluency Minor errors (50.59% and 46.86%), followed by Accuracy Critical errors (22.93% and 30.92%), Accuracy Minor (18.55% and 17.57%), and Fluency Critical errors (7.93% and 4.65%) (65). In terms of the impact of translation briefs and search conditions on PE quality, the data obtained from the group with inexperienced translators did not exhibit any significant difference regarding all the error labels. In addition, there was no significant difference when it came to the control of search conditions regarding all the error labels. The authors interpret this outcome as the result of the disparity in translation competence. Because inexperienced translators generally have less professional competence in research skills compared to experienced translators, they had significantly less Accuracy errors in post-editing when using the prepared set of terminology than when asked to search online even though the research capabilities were limited (69). In other words, participants with less experience seem to be able to make better use of the termbase than free search but are largely indifferent to the translation brief (67).

Another study which presents rare results on the impact of external resources during post-editing was conducted by Daems, Carl et al. (2016) where the researchers analyzed the data obtained from 40 translate-from-scratch and 40 post-editing sessions of ten graduate-

level students' translations. They used readability metrics to assess the corpus of texts to be used and in the final round had eight newspaper articles of 150–160 words, each consisting of 7–10 sentences (114). The experiment took place in two sessions. In the first, the procedure was as follows: Participants filled out an introductory survey about their experience with a focus on post-editing; in the second step, they performed the LexTALE test⁸² to measure their English proficiency; after that, in the third step, they copied a text of 150 words to become familiar with the specially-equipped keyboard and the chin rest of the eye-tracker, then they translated a text of four segments and post-edited the same amount of text in the CASMACAT interface and in the final round they translated two more texts and post-edited another two (116). In the second session, they took the same steps of getting familiar with CASMACAT interface with four segments to be post-edited and another four to be translated manually. After this initial warming up phase, participants completed the actual task where they had two texts to be translated manually and two texts to be post-edited. After this, they had to look at the texts again and highlight “the most problematic passages” for one translation task and one post-editing task and were asked to add comments to these in a Word document. At the end of the session, participants had to fill out a post-experiment survey, asking them about their experience and their attitude towards post-editing.

The results indicated that the ten translation students spent more time in external resources when translating than when post-editing even if the time spent in each specific resource was not significantly different between the two (121). In terms of translation quality, the results did not “seem to be significantly influenced by one specific type of resource, but rather by the overall time spent in external resources, as well as by the task type” (131). However, there was a significant difference in error scores between post-editing and translation from scratch: in the former task, longer consultation of external resources was accompanied by higher overall error scores; on the other hand, the opposite was true for translation from scratch where longer consultation of external resources was accompanied by lower overall error scores. The authors suggest these observations lead them to think “that

⁸² For more detailed information on the hypotheses behind modeling a valid lexical test for learners of English, see Lemhöfer and Broersma 2012.

participants are more successful in problem solving by consulting different resources when translating than when post-editing”, supporting the argument that post-editing requires different skills from human translation (see Yamada 2015) (131). Additionally, another result they suggest is that extra time spent consulting dictionaries in fact brings about an increase in acceptability, perhaps compensating for the loss in productivity (131). Finally, in the case of the impact of each type of external resource on adequacy quality, the results indicate that spending more time in encyclopedias does not bring about a decrease in error score, but rather an increase, suggesting that longer searches do not necessarily result in better translations with regards to adequacy (131).

The most comprehensive study in terms of the priming effect during the translation process, Bangalore et al.’s study on syntactic variance and priming analyzes the syntactic entropy effects observed in translation condition using German and Spanish datasets found in the CRITT-TPR database with sixteen subjects (2016). Assuming that both the ST and TT language are simultaneously activated during the entire translation process, the translator is assigned to explore and select potential target text elements as (s)he starts reading the ST sentence. Because both languages are assumed activated to the same degree, the probability that they have an impact on one another is very high and this interaction of ‘mutual’ influence is briefly defined as priming (216). The main question of this process research attempts to determine “[a]t what point during the translation process does the translator start with the mental production of the target text, and to what extent does this mental production process interfere with source text comprehension” (215). The authors stress that the concept of “literality” in TS and criteria associated with it proposed by Carl and Schaeffer⁸³ (2017), is particularly in line with research on translation entropy. They state that “the translations are structurally literal if an ST sentence is translated into the target language with a single syntactic structure by all translators in a given sample” and thus, explain that syntactic

⁸³ According to Carl and Schaeffer’s definition, a translation is literal if (1) word order is identical in the ST and TT, (2) ST and TT items correspond one-to-one, and (3) each ST word has only one possible translated form in a given context (2017, 217).

entropy is, in fact, an indicator for the literality of translations on a syntactic level (217). If priming as a psychological effect is going to be measured, the questions related to units of measurement, its causes, and its impact on cognitive effort during the translation process, and whether priming effects “modulate” the cognitive effort need to be answered. On that account, the authors formulate these two hypotheses:

(H1) • Segments with low entropy values reflect priming effects and are highly correlated with lower behavioural measures as compared to segments with high entropy.

(H2) • We predict that priming probability has a negative effect on behavioural measures such that items which are highly likely to have the same syntactic structure as the source sentence receive less attention than those sentences which are highly unlikely to have the same syntactic structure as the source sentence. It is expected that priming probability interacts with syntactic entropy.” (Bangalore et al. 2016, 219)

In order to test these hypotheses, the authors model two tasks for the experimental setup, first a translation task between one source language and three different target languages and second a monolingual copying task where participants were asked to read the same STs used in the translation task and type them. The copying task apparently serves as the control condition where if syntactic entropy effects are observed it would mean they also “represent monolingual source-language-related processes” (219-220). If otherwise, it is very likely that these effects are solely relate to target-language-related processes. The authors obtained both the gazing data and keystroke-logging data. The results indicate higher behavioural measures, i.e., total reading time and fixation durations can be considered an indication of “competition between multiple syntactic translation equivalents and the selection pressure generated from a set of co-activated syntactic realizations increasing cognitive load” as syntactic entropy was not a significant predictor of behavioral measures in the second copying task. These results confirm the first hypothesis that explains co-activation of linguistic systems during translation (235).

In the case of syntactic structural repetition, it was observed mainly in the vicinity of low syntactic variation in the translation task, which indicates that many translators were

structurally primed by the source. Therefore, these results also confirm the first hypothesis. As for the second hypothesis, no significant facilitation effect was observed for the interaction effect of syntactic entropy and priming probability (H2). In other words, priming did not necessarily have the expected degree of influence on syntactic variation, leading the authors to suggest that other processes that regulate variation could have been underestimated. For instance, if primed structure does not comply with target norms, translators can choose another alternative that is compatible with the norms and if more than one translator chooses the same structure, the resulting entropy value would be lowered to a priming effect and yet would not display a facilitation effect. Therefore, the interaction between priming probability and syntactic entropy did not behave as expected and there might be some other causes/variables effecting it (236).

The literature on the priming effect is not large when it comes to using MT as a pretest. Yet two other studies illuminate the way for further research. In a preliminary work by Natália Resende et al., the objective was to investigate the influence of Google Translate's MT system on the mental processing of English as L2 (2020, 245). Therefore, they conducted a syntactic priming experiment⁸⁴ using Google NMT output in the pretest design where participants' linguistic behavior before and after Google Translate was compared. This study also expanded on previous research as it is the first one to test syntactic priming from text (comprehension) to speech (production) by means of a cross-linguistic task (translation task) using software. The authors hypothesized that the user's syntactic choices when producing speech in a second language will mirror the syntax of the MT output when MT was involved in the translation process (248). The experiment asked the participants to take part in a game involving two stages, a pretest and a priming test phase. In the pretest phase, participants were asked to translate sentences describing images from Portuguese into English using words provided below each image. The pretest had 26 trials, 20 of which were image descriptions with a noun phrase subject, an auxiliary verb and a complement and 6 of which were filler trials with a subject (definite article + noun), an auxiliary verb and a complement.

⁸⁴ The authors refer to Bock's original experiment on priming effect (1986) and explain that it is also called "inverse preference effect" which states that the most uncommon structures drive the priming effect hypothetically (Heyselaar et al., 2017a; Heyselaar et al., 2017b).

In the priming test phase, after a translation task using Google Translate output, they had to describe images displayed on a screen aloud (249). The priming phase also had 26 trials and both phases were self-paced. The participants' formulations were either coded "1" if they produced NP constructions mirroring the Google Translate output and "0" if they produced other syntactic combinations (250). The authors also measured English proficiency as a variable of the study's hypothesis. The results show, as expected, participants used significantly more NP constructions, hence mirroring the syntactic structure previously seen in the Google Translate output when speaking in English compared to their descriptions in the pretest phase, most of which had PNP constructions (251). The results also show that participants at higher levels of English proficiency produced more NP constructions during the priming test compared to participants at lower levels of English proficiency. The authors think that based on previous research this is related to participants' attention being focused more on resolving semantic ambiguities than on parsing problems if their proficiency level is low. On the other hand, if participants' proficiency level is high, they tend to be more sensitive to syntax (252). Thus, the results indicate that this syntactic priming effect is modulated by English proficiency levels.

In a more recent study, Resende further questions whether the interaction between English language learners and Google NMT system could result in learning and increased production of a challenging syntactic structure in English that differs in word order between speakers' L1 and L2 (2021). The same experiment setup and procedure was used. Additionally, the study also measured the syntactic priming cumulative effects. The results show a robust syntactic priming effect as well as a robust cumulative effect and indicate that GNMT not only facilitates learners' awareness of the differences in word order differences between L1 and the target language, but also influences students to "unconsciously" produce sentences in the target language using the structure seen in the MT output. In terms of the cumulative priming effect, increasing the subjects' experience with a certain structure affects the magnitude of priming (2021). These studies on priming demonstrate some very clear results when it comes to second-language learning. Yet how effective MT is in priming when

the main variable of the study is related to post-editing performance still needs to be questioned.

Apart from these studies on priming effect, which are not necessarily concerned with post-editing performance, Ana Guerberof-Arenas conducted an experiment for correlating productivity and professional experience in a post-editing task setup which also involved fuzzy matches obtained from TMs (2014). This study repeats almost the same procedure she followed in an earlier one (2008) where 9 professional translators —8 translators and 1 reviewer— were asked to post-edit SMT output, translate from scratch in an online tool and also to complete a post-task questionnaire with 17 questions about their experience in localization, tools, subject matter and post-editing MT (13). The output they edited comprised MT (262 words) and TM (264 words produced using Trados to create fuzzy matches in 80-90 percent range) and they were also asked to translate from scratch in the “New” category (265 words) (12). The results demonstrate that in terms of productivity gain, translators performed better if they used a translation aid and the gain was higher in MT segments than in TM segments; 25 and 11 percent respectively. Yet, the standard deviation was extremely high and the range of variation was very pronounced (14). In terms of processing speed, translators processed more words per minute in MT than in TM or New segments and more words in TM than in New segments on average. But again, the standard deviation was extremely high (14). When it comes error analysis, according to the reviewer there were 17 changes needed in the text and these comprised text modifications, insertions and deletions, meaning there were 5 existing errors and 17 changes to make in the TM segments and 27 in the MT segments overall (15). However, the authors observed that in all eight cases there were more errors in the TM segments than in any other category annotated by the translators while the errors for the New and MT segments varied among the subjects (15-16). The authors interpret these results to have been caused by a common tendency among translators when they are presented with a text that “flows naturally” like a human translation, to seemingly pay less attention to how accurate the sentence is. Yet since errors in MT segments are so obviously wrong, the mistakes seem to be easier to detect and also because correcting the errors in the TM required the translator to look at the source text and

not just focus on the proposed target might have caused these segments to have more errors. But, because using MT helps translators clarify difficult aspects of the source texts and therefore improve general comprehension of the text, there were few errors annotated in those segments (16).

When it comes to types of errors, only 2 were found in the New segments as opposed to 9 in both MT and TM segments from the overall 20 Terminology errors. This can be interpreted as a result of translators' behavior in consulting the existing glossaries more when they are presented with new texts, rather than questioning the proposed terminology used in MT and TM (16). Therefore, in the case of an experimental setup instructions should be provided to reviewers or translators to specifically check glossaries before the translation process begins (16). Even when errors were considered, using MT was still more productive than no aid at all, which contradicting the study's first hypothesis, which stated "*the time invested in post-editing one string of machine translated text will correspond to the same time invested in editing a fuzzy matched string located in the 80-90 percent range*" (17). Interestingly, translators with less experience seem to make better use of both translation aids than those with more experience. The most important aspect of this study's results is that the translators with no experience had very similar processing speeds for both MT and TM segments (18). The author states that it indicates "translators with more experience are more accustomed to MT output and this familiarity prevents them from seeing very visible errors precisely due to this familiarisation" (19). This result is clearly related to more experienced translators being primed by MT output beforehand. In the case of the second hypothesis, which stated "*the final quality of the target segments translated using MT is not different to the final quality of New or TM segments*" (12) the results show that the quality produced by the translators is notably different when they use no aid, MT or TM, even if the number of errors found in MT segments is closer to those found in New segments (20). The third hypothesis which stated "*the greater the technical experience of the translator, the greater the productivity in post-editing MT and TM segments*" (12) was, on the other hand, partially proven. In a more recent study, where 24 translators and three reviewers were tested to assess productivity, words per minute, and quality data, and errors in final target texts, in the post-

editing of machine translation (MT) and fuzzy match segments (in the 85 to 94 range) for focusing on “levelling effect”, Guerberof Arenas had one major and one minor hypotheses:

[T]he greater the experience of the translator, the greater the productivity in post- editing MT match and Fuzzy match segments [...] [T]his technical experience will not have an impact on the quality (measured in number of errors) as was observed in the pilot project (Guerberof 2008) (3).

Results of the post-editing experiment data and the corpus of texts post-edited here again involving MT output, fuzzy matches in the 85 to 94 range and No matches, were grouped into clusters in order to correlate experience with speed based on number of words processed per minute and experience with quality by measuring the errors marked by the three reviewers in each match category (1). Both studies (Guerberof Arenas 2008 and 2014) reviewed here employed the LISA QA model. Half of the translators in this study had 8 or more years of experience in localization industry and more than 60 percent had experience in post-editing according to answers given for the questionnaire (5-7). The results indicate that in terms of processing speed, the cluster with least experience took full advantage of MT matches (10) and clearly the slowest cluster had no experience in post-editing. Nonetheless, the clusters with highest experience were still the best performers in terms of speed, with more than 30 words processed per minute (11). However, when a linear regression model was applied to the descriptive data, there were no statistically significant differences between Clusters in the No match category, contrary to previous studies by De Almeida and O’Brien (2010) and Guerberof (2008), which showed that faster translators were also the more experienced ones (13). Yet, in terms of number of errors in the Fuzzy matches category, quite expectedly the fastest and least experienced Cluster had the highest number of errors, most of which were in the Terminology category (16), and the slowest Cluster had the least number of errors according to three reviewers, indicating that though slow, they were ‘thorough’ and ‘rigorous’ (15). It seems relatively novice translators’ data signal “a lack of attention to certain important aspects of the translation process” (16).

The author finds the results related to the number of errors in the MT matches category “particularly interesting” because the difference between clusters is not as pronounced as

with the Fuzzy matches, indicating a levelling effect in favor of novice translators (17). When it comes to the number of errors in the No match category, the results are more akin to those of the Fuzzy match than to the MT match results, with the novice translator Cluster clearly having the highest number of errors, and the other three clusters being very close in results (19). Overall, the levelling effect was confirmed for the least experienced Cluster (21), parallel to earlier studies by De Almeida and O'Brien (2010) and Guerberof (2008). This group of translators made more mistakes mainly because they did not follow instructions and hence avoided the glossary, resulting in a higher speed but poorer quality (22). Consequently, the author suggests that senior translators might have a different idea of quality than novice translators while novice translators are more tolerant of MT errors and structures (23). In that respect Masaru Yamada quite sensibly asks in the title of one of his two recent studies "Can college students work as post-editors?" (2014 and 2019).

The preliminary study by Yamada explores the possibility of employing college students as post-editors in real-life settings and therefore, assesses students' potential aptitude for PE, the final quality of student post-edited products, and students' relative qualifications as post-editors by correlating perceived ease of task, editing quantity for measuring textual similarity between raw MT output and final product, and quality of final product with students' course grades (50). The text to be post-edited, a Wikipedia article on Steve Jobs, was translated by Google Translate (SMT), no CAT tool was used and students post-edited this raw MT output by overwriting it in a Word file (54). 43 students were asked to full post-edit the 486-word English text to Japanese and the test was done as a take-home exam due in a week (54). Three professional translators separately completed the same PE task to measure the quality of students' post-edits (55). For the first correlation on perceived ease of task, the investigation found that over 74 % of students felt PE to be an easier task than HT, whereas 26 % did not. Students who did not find PE easier were deemed to be unqualified post-editors alongside students who received poor grades in a traditional translation course. But it must be noted that A-students were not always qualified post-editors (49). In the case of revision amount, roughly 61.5 % of the raw MT output survived the PE process among students, whereas professionals had to revise about 38.5 % of the MT output to raise the product quality to the

requested level. Therefore, the translators' perceived workload (ease of PE) was 73.3 on average, where the base value of 100 is the reference for a conventional HT task. In other words, the participants experienced a 26.7 % workload reduction on average by PE (56). However, in terms of PE quality, no students met the level of quality achieved by the professionals ("errors totaling 1 or fewer"), yet the highest quality translations were achieved by 7 students (16 %) with a PE quality score of -2 (56).

According to the answers given to the post-task questionnaire, for the question "Did you find PE easier than HT?" 32 out of 43 (74 %) students responded *yes* and 11 students (26 %) answered *no*. In their analysis on data from PE quality, when the threshold of Pass/Fail was lowered to 5 errors or fewer, only one participant (5 %) in the pass group answered *no* to this question while 10 (40 %) did in the fail group (58). For the question on rating PE effort, the total average of 75.1 apparently shows that the PE task resulted in approximately a 25 % reduction in translation effort (63). Still, correlating the PE quality with ease of effort, when students who found PE harder compared to HT was eliminated, the score for ease of PE for the remaining participants in the fail group was 61.3 on average and 66.1 for the pass group, meaning almost equal ease of effort and thus, no significant difference for both pass and fail groups (63). Obviously, students who do not consider PE to be an easier task than HT are probably not good post-editor candidates. Moreover, in the case of revision amount, students in the pass group made more revisions than those in the fail group (64). Overall, it can be summarized from the results that students who get low grades in a traditional translation class are less likely to be high-quality post-editors, while getting a high grade does not guarantee that A-students will be qualified post-editors. The results confirm the argument made for and reviewed about PE in Chapter 2 that PE and HT may require different skill sets and training (65).

Yamada's more recent study using the same experimental setup with Google NMT output also presents results from comparing the earlier study with the more recent one (2019). The participant profile was nearly identical with 28 students who were majoring in English, not in a translation-related discipline, and took the practical English course whose instructor was the author. They also had some prior knowledge about PE (92). Again, the study had three

main questions about the PE cognitive effort measured in the perceived ease of task (Q1), PE revising effort measured in the amount of editing⁸⁵ (Q2) and finally, the quality of the students' final post-edited product measured in the degree of textual similarity (Q3) (91). The hypotheses formulated from these questions were NMT+PE would decrease the students' post-editing effort for Q1, translators would make fewer revisions for Q2 – the closer the values are to 1 means the closer the final post-edited content is to the raw MT output – and perceived workload would be lower for Q3 (91). The ST was again the Wikipedia article on Steve Jobs, translated using Google Translate NMT from Japanese to English and was laid side-by-side in a Word file format. Participants were asked to post-edit it by overwriting it and answer the questions in the questionnaire about ease of PE (92).

For Q1, 20 out of 28 students (71%) responded yes to the questions “Did you find NMT+PE easier than HT?”, while 74% answered yes in the previous experiment with SMT+PE (92). Moreover, when asked to rate their PE effort with a numeric ratio, a total average of 79.1% suggested that the NMT+PE task resulted approximately in a 21% reduction in PE effort when compared with HT (92). These results, in Yamada's words, were found to be “rather unexpected” and the first hypothesis was not confirmed (93). In the case of amount of editing, the NMT groups average was 0.210 while the SMT group's average was 0.247, demonstrating that post-editors using NMT made fewer revisions than those in the SMT group (93). Therefore, the second hypothesis was confirmed. However, the amount of editing needed for NMT+PE was less than the amount of self-revisions, indicating that these results could be interpreted to mean that NMT+PE was more demanding as those fewer edits required more concentration than self-revision (93). In the case of the third hypothesis focusing on perceived workload, according to error annotation NMT output had a total of 27 errors, 10 of which were major, whereas SMT output had 43 errors in total, 31 of which were major (94). But as answers given for the first hypothesis does not support the assumption,

⁸⁵ Yamada bases this question on the study conducted by Comparin and Mendes (2017) where one pertinent finding was that 85% of the errors found in the raw MT were correctly revised through human post-editing. Yet, as the number of fluency errors were less, a relatively high number of accuracy errors were still not corrected (2019, 90).

the third hypothesis is still not confirmed (94). Finally, as for the quality of students, after comparing the number of corrected errors with the students' annotations⁸⁶, the study reveals that 68% of errors were corrected for NMT+PE whereas 77.7% of errors were corrected in SMT+PE and the variation in correction rate was almost the same (94). Yet, in contrast to the assumption about NMT, student post-editors of this second round of PE did not necessarily outperform student post-editors using SMT+PE and could not meet the professional quality standard of an 85% error correction rate (see Comparin and Mendes 2017). Nonetheless, the number of errors remaining in the final translation after PE decreased by nearly half compared to SMT, indicating a major achievement for NMT. Thus, it is apparent that NMT had a levelling effect on non-professional post-editors in this study as well (94), parallel to Guerberof Arenas's results (2014). After analyzing the error annotations and categories in detail, Yamada asserts that "NMT+PE requires almost the same competence as ordinary human translation. Therefore, translation training is necessary for students to be able to shift their attention to the right problems", i.e., mistranslation (102).

Overall, the studies reviewed here reveal what are called priming and levelling effects for different participant groups who take part in post-editing and quality evaluation experiments. Participants with divergent levels of experience tend to interact with MT output differently as their quality evaluations and processing times and steps present results which are statistically different, sometimes even contrasting compared to novice translators who are not proficient in translation skills such as knowing when to research terms and using research outcomes accurately and adequately. In this sense, their proficiency in translation is similar to L2 acquisition as there are many corresponding strands of arguments on priming from a psycholinguistics-based approach and post-editing quality evaluation from a Translation Studies-based approach. In the next chapter, the experimental setup for the present thesis follows a design which incorporates the necessary aspects of both approaches.

⁸⁶ For this part of the study only major errors were taken into consideration.

CHAPTER 5 A 2-TIERED EXPERIMENT ON PRIMING WITH SPECIALIZED TEXTS

This chapter presents the 2-tiered priming procedure employed for the post-editing study. By 2-tiered, we refer to participants' post-editing (Task 1) results and error annotations (Task 2) in the priming set, followed by a test set with both tasks. The crucial point in this procedure is that before Task 1, all participants were asked to complete a pre-experiment survey where they had questions concerning their experience in post-editing and translation and their opinions regarding the use of MT and their speculations about the requirements expected from translators related to prospective translation project management and service provision. Therefore, before delving into the details of this survey, the corpus of ST used for creating MT output is introduced first, to be followed by the details for preparing a priming set for post-editing similar to Bock's (1986) descriptions.

5.1 Sustainability reports and their stylometry

The decision to use parts from Henkel and Siemens Sustainability Reports was based on the literature on document-level quality evaluation as two parts selected from these reports were extracts which did not necessarily have more technical, expertise-requiring terms in them and dealt with general business relationships and these organizations' "most significant impacts on the economy, environment, and people" to enhance "transparency on the organization's impacts and increase organizational accountability (2023, GRI Standards). But why are these texts important in the case of technical reporting and post-editing research? There are a few answers to that question.

First of all, in the Official Gazette dated 29.12.2023, the Public Oversight, Accounting and Auditing Standards Authority⁸⁷ (POA) announced Turkish Sustainability

⁸⁷ Formerly known as Turkish Accounting Standards Board (TASB), POA was established in November 2011 "in order to remove fragmented structure in the field of accounting and to fulfil the commitment given to the EU by ensuring an effective auditing and public oversight system" ("Public Oversight" 2015, 12) and it is "*the sole supreme authority in determining auditing standards and ethics, authorizing independent auditors*

Reporting Standards which determine the principles to be followed in sustainability reports. Institutions, organizations and businesses which had values amounting those stated in the Decision and had a total of at least 250 employees were asked to produce sustainability reports following the Turkish Sustainability Reporting Standards and Board Decisions that were fully compatible with the global S1 and S2 sustainability reporting standards of the International Sustainability Standards Board (Ministry of Trade Green Deal ⁸⁸).

Secondly, though these texts are a product of the marketing/legal domain, they deal with the most generic issues related to the world of commerce, which anyone who received introductory level education on social sciences might have very easily some basic knowledge of. Also, because the participants were given freedom to search online, they could get a glimpse into the context and meaning of a few key words. The contents of the international standard reporting according to the Global Reporting Initiative is divided into three main categories: universal standards, sector standards and topic standards (GRI1: 2021, 6). Yet, the headings in a sustainability report are exemplified as General Provisions (TSRS S1), Climate-related Disclosures (TSRS S2), Disclosure of Relevant Financial Information (TSRS S1), Explanations on Social and Employee Rights, Environmental and Social Impacts and Management and Internal Control (“Egemenoglu”⁸⁹). Overall, the Turkish and the international standards for sustainability reporting correspond to one another. But it should not be forgotten that the two documents selected for translation, PE and revision for the present study are going to be translated and not re-written. Therefore, the participants are expected to follow the general structure of the text and do some research on the topic before the experiment as they are provided with the source text beforehand. After this concise

and audit companies under a public oversight system and monitoring their activities within the frame of quality assurance”, in the process of EU negotiations” (11).

⁸⁸ For more information on Green Deal, see the Ministry of Trade website: <https://ticaret.gov.tr/dis-iliskiler/yesil-mutabakat/duyurular/turkiye-surdurulebilirlik-raporlama-standartlari-ve-uygulama-kapsamina-iliskin-kurul-kararlari-resmi-gazetede-yayimlanmis-ve-belli-sartlari-saglayan-isletmeler-icin-surdurulebilirlik-raporlamasi-zorunlu-hale-getirilmistir> (Accessed February 16, 2024) and <https://www.trade.gov.tr/exports/green-action-plan> (Accessed April 13, 2025)

⁸⁹ For a summary of GRI standards, see: <https://www.egemenoglu.av.tr/en/news/turkish-sustainability-reporting-standards-tsrs-and-scope-of-application-of-tsrs-were-published-in-the-official-gazette> (Accessed May 1, 2025)

background knowledge on the topic, now the MT systems used and how the priming set was prepared is going to be discussed.

5.2 MT systems used and preparing the priming set

The Corpus of MT output to be post-edited has parts from two different multinational companies' sustainability reports. The parts selected do not necessarily include specific terms associated with sustainability and therefore can be considered general documentation in the marketing domain. At this point, it is crucial to note that after completing Task 2, a participant suggested that even if they are considered in the general domain, these marketing texts still require some unique expertise. Following this comment, the present study attempts to challenge the boundaries of sub-domain categories – in this case, general and marketing– (c.f. Castilho and O'Brien 2023⁹⁰ and employs Danielle Saunders' working definition that “[a] domain may be a well-defined topic, text of a specific provenance, text of unknown provenance with an identifiable vocabulary distribution, or language with some other stylometric feature” (2022, *EAMT Proceedings*, 9). This definition applies to the samples of this study very comfortably as the stylometric features of sustainability reports addressing customers with a corporate voice is quite visible. Some of the participants even suggested that they peruse through the corporate websites to understand their tone of voice and style for the PE task. Thus, MT systems could identify the domain using these features from a theoretical point of view.

The MT systems used in the present study are two – one black box and the other customizable with an easy-to-use interface. The NMT systems selected were Google Translate NMT and Kantan AI MT. Initially, we attempted to customize one of Kantan's NMT systems with the custom corpus compiled with article abstracts found on Turkish digital library for scientific journals called Dergipark in both English and Turkish. But because the quality of NMT output customized with this corpus resulted in very low-quality

⁹⁰ Castilho and O'Brien define domain as “any combination(s) of content, topic, genre, text type, register and style, depending on the purpose of the data, and on what is needed from a corpus” (University of Surrey Convergence 2023).

output as an evaluator pointed out, and considering that the output would not be used in the industry as it would not enhance translators' efficiency, we decided to continue using a baseline corpus provided by Kantan AI with the specifications described in the Table below:

Table 3: Automated Scores Distribution for Kantan Pretrained Engine

Automated Scores Distribution - Summary Report for Job #131406					
Engine Name	Source	Target	Word Count	Unique WC	Mono WC
2019-BASELINE-tr-en-NMT-Copy	tr	en	40376022	1478177	0
Interval Segments	F-Measure				71%
	<40%	40-49%	50-69%	70-89%	90-100%
	39	21	145	130	165
Interval Segments	BLEU				59%
	<40%	40-49%	50-69%	70-89%	90-100%
	169	37	98	42	154
Interval Segments	TER				41%
	<40%	40-49%	50-69%	70-89%	90-100%
	285	33	99	50	33
Interval Segments	Perplexity				7%
	<40%	40-49%	50-69%	70-89%	90-100%
	495	0	0	0	5

For Task 1, as the present study focuses on the priming effect, some of the output, if the sentence structure permitted it, was changed to have primers following Bock's priming experiment design. Where sentences which clearly required an active voice were changed to passive voice by altering the present tense suffixes such as the progressive suffix "-uyor" with perfect tense marker "-ulmaktadır" at the end of sentences (Primer 2) and the verb in the verb phrase "ilerlettiğini" with the transitive object was changed into "ilerletebileceğini" to make the translation read more literally with the abilitative marker "-(y)eBil" (Kornfilt 1997, 343), followed by the future tense marker "-(y)EcEk" (340) meaning being able to (Primer 6); instead of transitive and intransitive priming, we changed the position of NPs which were initially acting as the main verb (a noun clause) as in "gerektermektedir" and used them as prepositional noun clauses "gerekiren" (Primer 3). Also, in order to get a real in-depth understanding how the priming segments would work in PE, the postposition "için"[for] working as a subordinating conjunction was altered into "amaçlı" [for the

purposes of] in a subheading (Primer 1). Furthermore, the noun clause “lider bir teknoloji şirketi” [a leading technology company] which had the main verb position in the Google Translate NMT was changed into a subordinate noun clause in another segment (Primer 4). Finally, the verb phrase “artırmak” with the active voice was altered to the passive voice version “artırılmakta” and instead of being connected with the postposition “için” [for], they were connected with the conjunction “ve” [and], followed by a noun clause referring to these verb phrases through expressing the objective with “bunun” [these] in Primer 5. Examples from the priming set can be seen in the table below:

Table 4: Examples from Priming Segments

Primer 1	NMT output (NMTG)	Sürdürülebilirlik <u>için</u> güçlü bir çerçeve
	Priming Sentence (PS)	Sürdürülebilirlik <u>amaçlı</u> güçlü bir çerçeve
Primer 2	NMT output (NMTG)	Çerçeve, temel sürdürülebilirlik değerlerimize 360 derecelik bir yaklaşım <u>sunuyor</u> .
	Priming Sentence (PS)	Çerçeve ile temel sürdürülebilirlik değerlerimize 360 derecelik bir yaklaşım <u>sunulmaktadır</u> .
Primer 3	NMT output (NMTG)	Küresel iklim değişikliği, insanlığın bugün karşı karşıya olduğu en büyük zorluklardan biridir ve acil ve iddialı eylem <u>gerektirmektedir</u> .
	Priming Sentence (PS)	Küresel iklim değişikliği, acil ve iddialı eylem <u>gerektiren</u> , insanlığın bugün karşı karşıya olduğu en büyük zorluklardan biri.

Primer 4	NMT output (NMTG)	Siemens, endüstri, altyapı, mobilite ve sağlık hizmetlerinin dijital ve sürdürülebilir dönüşümünü yönlendirmek için tasarlanmış bir portföye <u>sahip lider bir teknoloji şirkettir.</u>
	Priming Sentence (PS)	Siemens <u>lider bir teknoloji şirketi</u> olarak endüstri, altyapı, mobilite ve sağlık hizmetlerinin dijital ve sürdürülebilir dönüşümünü yönlendirmek için tasarlanmış bir portföye <u>sahiptir.</u>
Primer 5	NMT output (NMTG)	Siemens'in endüstriyel metaverse ve dijital ikiz teknolojileri, üretim tasarımını optimize etmek, üretim verimliliğini <u>artırmak</u> ve CO2e emisyonlarını ve enerji, su ve hammadde tüketimini <u>azaltmak için</u> fiziksel fabrikalardan ve ürünlerden gelen verileri tüm yaşam döngüleri boyunca <u>entegre eder.</u>
	Priming Sentence (PS)	Siemens'in endüstriyel metaverse ve dijital ikiz teknolojileri ile, üretim tasarımı optimize edilmekte, üretim verimliliği <u>artırılmakta</u> ve CO2e emisyonları ve enerji, su ve hammadde tüketimi <u>azaltılmakta ve bunun için</u> fiziksel fabrikalardan ve ürünlerden gelen veriler tüm yaşam döngüleri boyunca <u>entegre edilmektedir.</u>
Primer 6	NMT output (NMTG)	Ürünlerimiz, hizmetlerimiz ve çözümlerimiz, teknolojinin sürdürülebilirliği nasıl <u>ilerlettiğini</u> ve içinde yaşamak istediğimiz dünyayı şekillendirmeye nasıl yardımcı olduğunu göstermektedir.
	Priming Sentence (PS)	Ürünlerimiz, hizmetlerimiz ve çözümlerimiz, teknolojinin sürdürülebilirliği nasıl <u>ilerletebileceğini</u> ve içinde yaşamak istediğimiz dünyayı şekillendirmeye nasıl yardımcı olduğunu göstermektedir.

The ST, the MT output with priming segments and the Kantan MT output used in Task 2 were analyzed with Sketch Engine and the results below were obtained about this small corpus of text:

Table 5: Sketch Engine Analysis on ST, MT-Priming Set and MT-Test Set

	ST	MT-Priming Set	MT-Test Set
Tokens	673	605	568
Words	593	518	488
Sentences	24	24	24
Number of unique word forms including non-words	315	364	349
Tag	37	64	71
Lempos ⁹¹	285	277	275
Pos	9		
Lemma	275	250	

As seen in the table above, while the number of sentences remain the same in the priming and the test set, there is a significant change in terms of tokens, a minor change in the number of words, a very slight change in the number of unique word forms and lempos. Most studies in the literature review presented in the previous chapters tended not to analyze their corpus in detail using NLP techniques apart from Daems et al.'s (2016) study which uses newspaper articles with the highest possible Lexile⁹² levels (between 1160 L and 1190 L) (114). Because the materials for the PE tasks are not necessarily designed for education purpose, the analysis on the ST of the present study done with the Readable App revealed some expected results for a text on the margins of technical/non-technical texts with a low Flesch-Kincaid grade level (11.6) and Gunning Fox index (13.5), and with issues such as very difficult readability, very long sentence and hard work. Yet, it must be noted that most participants in this study

⁹¹ Lempos is defined as “a positional attribute, i.e. an attribute assigned to each token in the corpus” (Sketch Engine). The Sketch Engine website also states that “[i]t is a combination of lemma and part of speech (pos) consisting of the lemma, hyphen and a one-letter abbreviation of the part of speech, eg. *go-v*, *house-n*” (<https://www.sketchengine.eu/glossary/lempos/>).

⁹² The Lexile Framework for Reading provides educators with a measure of students' reading ability from beginner to advanced and is used for predicting how well a student will likely comprehend the text (“Lexile”).

had their bachelors' degrees in translation studies and at least some experience in translation at least. Therefore, the measurements for elementary and high school students' comprehension do not necessarily accommodate the expected comprehension level for this research.

5.3 The survey results

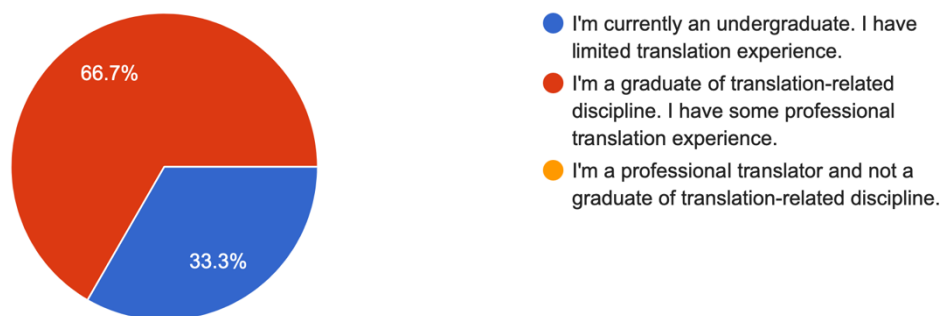
Before Task 1, all participants were asked to complete a pre-experiment survey consisting of 13 close and open-ended questions asking about their perceptions, expectations and experience of MT. The number of participants who volunteered for the experiment was initially 12 and yet not all completed the whole procedure. Some participants devoted extra time to sending their further opinions as additional information after taking the survey. The survey questions and the answers given are analyzed below.

Question 1 asked “Could you give information on the department you graduated from and your professional translation experience?” The choices can be seen in the image below. 66.7% of participants answered that they had graduated from a translation-related discipline and had ‘some’ professional, while 33.3% said they were studying mostly for a diploma in Translation Studies at the undergraduate level. In order to go deeper into their experience levels, Question 2 (Q2) asked the participants to specify a level between limited, 1-3 years and 3+ years. All respondents (12 people) answered the question as they did Q1. 58.3% of participants stated that they had limited experience whereas 41.7% opted for the 3+ years’ experience option. The specific choices for Q2 were as follows: “I’m currently an undergraduate. I have limited translation experience”, “I’m a graduate of a translation-related discipline. I have some professional translation experience”, and “I’m a professional translator and not a graduate of translation-related discipline”.

Figure 4: Answers given for Q2

Could you give information on the department you graduated from and your professional translation experience?

12 responses



Question 3 was about participants' previous post-editing and revision experience, and when asked a Yes/No question, and all the participants (12 of them) answered. 66.7% of respondents stated 'Yes', whereas 33.3% said 'No'. Question 4 was about participants' knowledge of post-editing and revision tasks and they were asked to explain their source of information. 8 respondents said "Yes" and 3 specifically wrote "No", expressing that they "have not heard anything" about these tasks. When it comes to their source of information, some pointed to lectures and professional environments and one suggested "some of [their] school tasks require post-editing", another one stated that for the course Quality Standards in Translation they wrote a final paper the subject of which was post-editing and MT, another one expressed that they had been doing it [PE and revision] to make the text more professional while one participant wrote that they had been post-editing professionally for over 3 years. Only 4 of the respondents provided their own definition of the processes, which are quoted below:

Post-editing consists of processes in which machine translation output is edited based on such criteria as accuracy and readability. Revision, on the other hand, has a more comprehensive scope, not only in-textual, but also extra-textual elements[sic].

The term "Revision" is commonly used to refer to the reviewing and correction of human translation in all respects, while post-editing is commonly used to denote the selective (light) or comprehensive (full) correction of machine translation output.

Post-editing is done via conducting a bilingual check on a machine translation output and editing it where needed. Revision is almost the same process, however the output doesn't necessarily have to be a machine translation, it can also be human translation.

[R]evision and post-editing mean to check the final product to see if it fits the agreed purpose of the translation.

Therefore, it can be argued that over 75% of all respondents knew the essential qualities of PE and revision processes as defined in the international standard. For Question 5 asking "If you have any (professional) post-editing experience, could you please give some details about the process of getting clients and the task? How often do you get post-editing jobs?", only 7 people provided an answer with one writing "NA", and two others "I don't have a job at the moment," and "No, I do not have". However, 4 others explained the processes in which they found clients as below:

I regularly get post-editing jobs from LSPs.

I have a single client which is a multilingual vendor (MLV) in the language industry. I regularly (everyday) get jobs from that client. About 40% of the jobs are labelled "translation", another 40% revision and the remaining 20% has miscellaneous labels. Those jobs that are labelled "translation" can be considered "full post-editing" jobs in that the regular procedure is to work, using a CAT tool, on the output of a customized MT solution. The output (MT) is visible simultaneously with the source segment, and the "translator" works on the output.

I only did post-editing for my article in one of my classes. I chose a blog and there I had the "review" section for a product first translated to artificial intelligence. I noted the mistakes made. Then I did post-editing myself. I chose a review consisting of 15 paragraphs. I noted the mistakes made by the artificial intelligence in each paragraph and I identified the mistakes according to the article made by Öner (2021). I did not do any other work.

I work as a freelance linguist. Currently, I'm working with the user experience department of a corporate firm manufacturing consumer electronics. Post-editing their user manuals is one of the main tasks that I perform as part of my job -aside from pre-editing and text optimization for AI. I have been hired solely for this purpose. Every time a new content is created or updated, I post-

edit it. The frequency depends on how often they create/update a content, which usually occurs every week.

As for Question 6 asking “If you have no previous post-editing experience, have you used publicly-available machine translation (MT) systems in professional translation jobs?”, all 12 participants answered, and 58.3 percent answered “No”, 41.7 percent “Yes”. When asked to explain for what purpose and how they used publicly-available machine translation systems in Question 7, we received 6 responses, copied below:

I used multiple MT platforms to translate a scientific book. I copied and pasted one sentence each time and performed the translation task using and editing the best MT output.

I never use public MT systems for a professional task

I have mostly used Google translation, since the language I have done translation from and to was Japanese and Google was successful at giving the corresponding terms for some Japanese characters that I had difficulty at telling apart, and helped me shorten the time that I would have lost if I were to look each character up.

I used MT systems for some of my schoolwork and my individual works. I tried to use it on novel translations by using simple steps of any simple MT system.

I answer this question because the previous question is mandatory. Even though I have previous post-editing experience, I think I can still answer the question. Yes, I have used. For the purposes of testing and training, I need to use chatbots for pre-editing and post-editing. However, I make sure that the content is not proprietary or confidential. Other than that, the content that I post-edit is translated by one of the MT providers (e.g. Google Translate, Yandex Translate, OpenAI etc.) which is automatically suggested by a localization automation platform that my customer asks me to use.

I have used machine translation for post-editing of a text.

For Question 8 “In line with the improvements in machine translation technology in recent years, did you observe any change in MT output?”, we received 12 responses, with 91.7 percent opting for “Yes” and the remaining 8.3 percent for “No”. For Question 9, asking for a specific example of the change mentioned in Q8, we received 11 answers as quoted below:

Statistical machine translation tools used to translate a text in its literal meaning, but I occasionally experience neural ones, such as DeepL, use even idioms depending on the context.

More accurate outputs in any kind of document trans. Accurate transformation on gender-related words, providing contextual alternatives and post-editing mode after trans.

Better outputs as they can understand the context better. (cultural nuances, providing options, style and tone etc.)

MT output has gotten more fluent and elaborate in word choice. Also, adaptive MT engines learn from corrections better.

recently, mt systems achieve more professional translations and this improvement is getting increase [sic]

The customized MT solution that I do -and are supposed to- use in my client's projects learn and imitate our terminology and phraseology.

It might be a far too easy answer, however, when I first started my translation training, the machine translation output of "Love is a four letter word" in Turkish was "Aşk dört harfli bir kelimedir", but now, maybe due to its manual training or neural training, the machine can translate it as "Aşk üç harfli bir kelimedir".

Some MTs offer translations that may be better than others. But they cannot provide the full meaning on a text-by-text basis. They give more appropriate suggestions than before, but they cannot capture the meaning we want. Also, the grammar is still very lacking. For example, I participated in the short story competition organized by the "Atatürk Kültür Merkezi Başkanlığı". And we were asked to translate Sait Faik Abasıyanık's short stories. After my own translation, I asked the AI to translate the same part. I realized that the AI cannot distinguish between subjects. In fact, it did the same thing when I gave it information about the story before asking it to translate. Faster but the meaning is not good enough.

I don't know if this is a good example, but in my research techniques course in 2nd year, my research topic was "The Usage of Online Translation Tools in Translation Students.". In the questionnaire, I asked the students about their experiences. One student shared a memory he had with a translation tool. "For example, a translation tool translated the phrase 'zaman aralığı' as time December.". This is the student's own statement.

MT applications have become more interactive (e.g. one can select one of the many equivalents of a source term or phrase within the same application easily or ask for rephrasing with a simple click). Also, being able to upload a whole document results in more context-aware translations, achieving better accuracy in terms of terminology. With generative AI tools or APIs and chatbots, which can be used as MT tools, one can use MT not only for translation but also for any other translation-related task such as pre-editing, revision, review, or proofreading through simple prompts or customized GPTs.

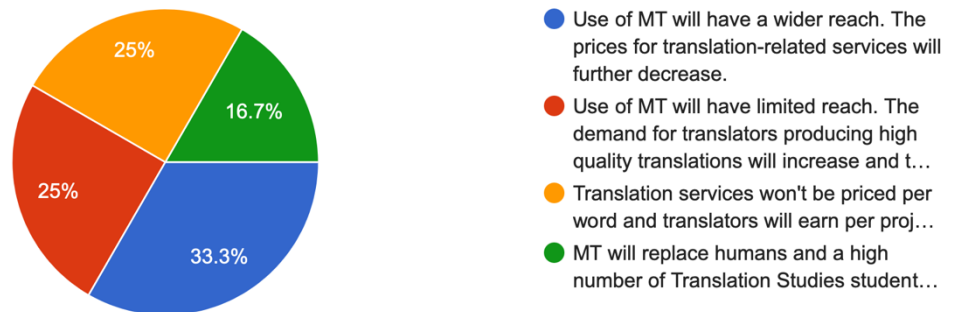
I cannot remember a specific example but I can say that the accuracy and the context correctness of translation improved a lot.

Question 10 asked “What do you think about the impact of machine translation in your future professional life?” and provided 4 answers based on the literature review of the previous chapters (C1: Use of MT will have a wider reach. The prices for translation-related services will further decrease. C2: Use of MT will have limited reach. The demand for translators producing high quality translations will increase and their services will be valued at higher prices. C3: Translation services won't be priced per word and translators will earn per project with services besides text production. C4: MT will replace humans and a high number of Translation Studies students won't be able find jobs.) Quite interestingly, one respondent suggested she would choose all the answers if she was given the possibility.

Figure 5: Answers given for Q10

What do you think about the impact of machine translation in your future professional life?

12 responses

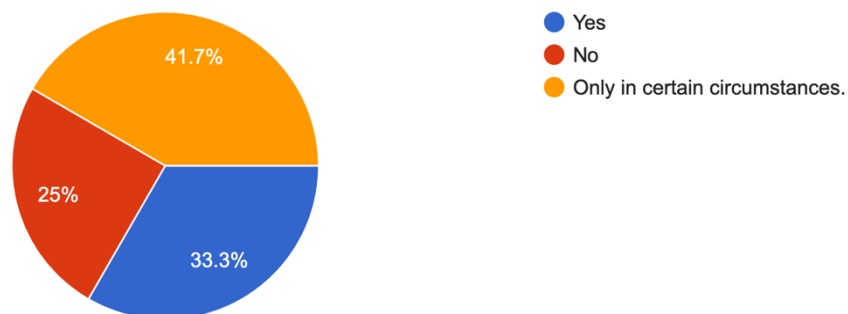


Question 11 asked “Do you find MT output reliable and safe for your daily needs and use it?”, we received 12 responses.

Figure 6: Answers given for Q11

Do you find MT output reliable and safe for your daily needs and use it?

12 responses



Close to the end of the survey, Question 12 asked “Specifically, in which circumstances do you find MT reliable?” and we received a mere 6 responses. Respondents interpreted “reliable” in varying ways in terms of the shelf-life of translation and its purpose. One said “For communicational purposes, to make complex language usage clear”, another wrote “Reliable for the texts without personal and safe data”, again another one focused on technical context “In some highly technical contexts, or in everyday speeches that are established and don't really have much room for errors”. Yet, someone stated an opposing view – as (s)he put it, “It can give accurate results on a word basis. But according to their fields. It can work for daily used words, but I do not find it safe for words in the field of medicine, law, culture or industry”. Another one indicated the shelf-life of the content and wrote “Daily use, not for literal or confidential documents”[sic]. One respondent clarified their point of view: “Even though I would not describe it "reliable", I find it safe to use MT for uncritical use cases or cases that are not legally binding or may have consequences for health or environment. Examples include getting the gist of a user's comment on social media or product reviews, automatic translation of subtitles or closed captions for social media content or news created in a second language that I don't speak. As long as I am aware that the output is machine-translated and the purpose falls under the categories that I mentioned earlier, I can always remain cautious against potential risks of MT while using it at the same

time.” For the final Question (Q13), asking “Do you have any experience in machine translation evaluation using MateCAT's quality framework?”, 72.7% percent of participants responded with a ‘Yes’ and 27.3% responded with a ‘No’.

Similar to the present study’s questionnaire, Öner and Öner-Bulut’s study⁹³ (2021) asked “If you think that the error annotation and post-editing experience you have gained during the study will help you (considerably) in translating from scratch experience, in what way or ways will this experience help you?” to the participants, who were MA Program students taking the translation Quality Standards course. The answers given by the respondents also provide further implications for the priming effect in an inconspicuous way as some respondents were not “sure about the potential contribution of error annotation and post-editing experience to translation from scratch”, and “emphasized that the experience would rather help them in ‘post-editing’ and ‘editing’” (112). Ironically enough, when asked “If you think that the error annotation and post-editing experience you have gained during the study will help you (considerably) in your future endeavors as a post-editor, in what way or ways will this experience help you?”, an overwhelming majority of the participants asserted that “the experience they gained would help them in their future roles as post-editors” (112). Therefore, it can be argued that the tasks of post-editing and revision are more similar in this group’s minds, yet they do not recognize that the skills they acquire for post-editing would also contribute to translation from scratch. In that case, if post-editing replaces translation in the MT workflow then it could be argued that post-editing can work as an augmented tool or mitigator between translation and revision in the conventional workflow. But would it be help for make the less experienced translators observe the results of priming they experience using fuzzy matches or MT output? That is a problem we attempt to lay bare in the next part.

5.4 Iterative text revision and editing

⁹³ Öner Bulut and Alimen’s study on human translator educator also provide answers to similar questions, see Öner Bulut and Alimen 2023. Çetiner 2022, 2021 and 2019 conduct research surveying undergraduate level students’ attitudes towards MT use and the sustainability of the profession.

The details for the experimental setup, the participant profile and the hypotheses of the present study can be found below.

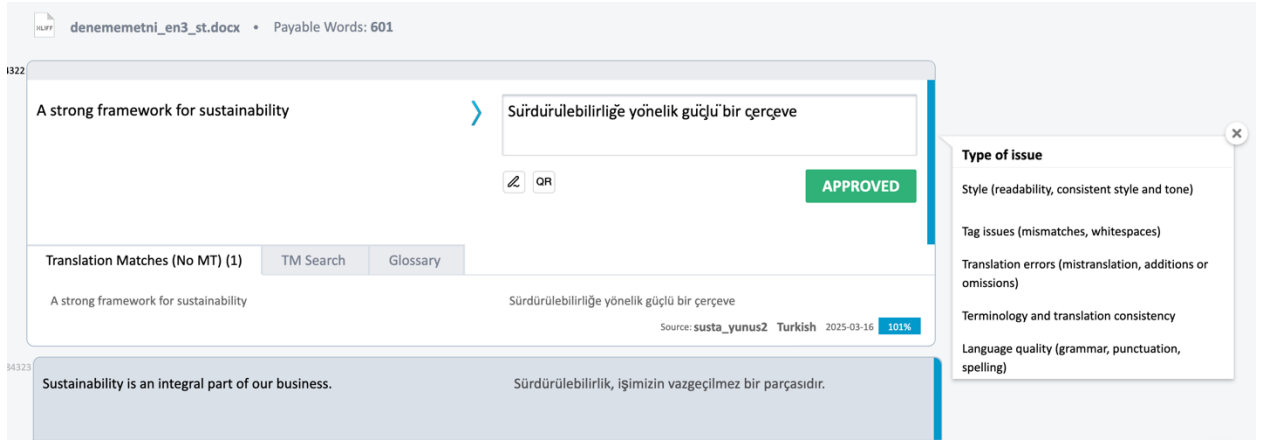
5.4.1 Participants

Participants who registered for the experiment had divergent levels of experience in translation and graduated mostly from Translation Studies-related disciplines or departments of Teaching English as a Second Language. Undergraduate students from Konya Karatay University – some people from this group only completed the pre-experiment questionnaire – graduate level (MA and PhD) students from various universities and lastly, 2 assistant professors working at Konya Karatay University and Yeditepe University volunteered in both the pre-experiment questionnaire, prepared in Google Docs, on and the following two PE tasks. The independent variable will consist of two levels: (1) the novice translators who had limited PE-related experience (the experimental group); and (2) the experienced translators with PE and revision-related experience (the control group).

5.4.2 Experimental layout

Both tasks were prepared on Matecat by creating a separate resource – a translation memory – for each task including Task1 and 2 for the same participant and not using the ModernMT Lite option provided by Matecat. The working interface for the participants in the experiment is shown in Figure 8.

Figure 7: Working Interface in Matecat for the Participants



5.4.3 The hypotheses

The four hypotheses of the present study are:

- If the translation phase in the workflow is replaced by MT, MT as an aid primes the novice translator in terms of speed. So, *the post-editing by less experienced translators will be faster than that of the professional translators who would be more attuned to errors and would need a longer period of post-editing time. Therefore, the number of total errors will be higher in the test set than the priming set.*
- After completing the priming task, translators would develop a familiarity with certain error types. As a result, because they think they are post-editing the same material in the test set, as some participants clearly suggested in conversation about the study after completing both the tasks, they might be less alert to the primed error categories and show a positive priming effect. Therefore, *the number of errors annotated in the test set will be less in the issue types annotated in the priming set.*
- If the overall quality of the MT output in the test set is found to be lower than in the priming set based on the evaluations, *translators will spend more time on post-editing in Task 2 and will not be primed by the output used in Task1.* This result would also demonstrate that priming to certain error categories is text-bound, meaning the

priming effect might be more evident in a document-level evaluation based on the participant's macro level estimation of the overall document at hand.

Before interpreting the results obtained from the experimental study, a trail-blazing study reporting on the outcomes of a pilot program by Inoue and Candlin (2015) using Task-Based Learning, which aims to fill the gaps between what novices have acquired through their training programs and professional requirements, will be briefly outlined. The authors sought to “formulate an alternative research-based pedagogical framework by drawing on a Task-Based Learning (TBL) approach to translator education and a holistic view of translation expertise” (59). Referring to Kirally's conceptualization of translation competence and ‘translator’ competence⁹⁴, key characteristics of TBL are as follows: “Tasks are of a holistic nature”, “Tasks are interconnected”, “Learning takes place in a professional environment”, and “Shift in teacher and student roles” (61-62). The participants of the study were six postgraduate students majoring in translation and interpreting in Australian universities and both the process and product of five tasks were evaluated by two professional translators followed by a set of interviews for identify learner perception on the usefulness of TBL (58).

Drawing on the social constructivist approach, the authors attempt to bring “authenticity” into translator education and training for a holistic view of translation expertise and state that the key characteristics of TBL related to their task design, including “tasks are of a holistic nature”, “tasks are interconnected” and “learning takes place in a professional environment” and “shift in teacher and student roles” were determined (61-62). For this reason, Willis' task cycle model was applied, as it provided “novice participants with a ‘bird's eye view’ of translation expertise in which key challenges⁹⁵” (62) were interconnected so that they would “enhance and scaffold” (63). Therefore, the prioritized conditions for the set of instructional tasks included that task components should be “designed to raise learner awareness of market demands and requirements” and that “learner autonomy is achieved through learners' and teachers' roles as active and collaborative agents and facilitators, respectively” (64). In order

⁹⁴ See Chapter 2.

⁹⁵ The authors explain that among these key challenges were “gaining an adequate understanding of a ST, searching for necessary information, building a sound relationship with a client, and achieving appropriate readability in accordance with a set target readership” (64).

to test their design, the authors collected four sets of data: whole class discussions recorded and transcribed, oral presentations by novice groups on key challenges, TTs produced by novice groups and finally, professional evaluations on novice groups' processes and products (66). The authors then analyzed the data obtained for examining how the professionals perceived learner performance in terms of translation processes and products and compared the results between sessions so they could assess the degree to which the learners developed techniques to identify and address translation challenges (67). The results indicate that while the cohort Group A "demonstrated an ability to identify where they needed to make improvements in their translating processes and products", and "were subsequently able promptly to rectify such weak points", Group B needed a longer period of time to rectify their weaknesses due to different "previous exposure to a learning environment that incorporated real-life requirements and collaborative learning experiences" (68). The authors also highlight that the quality and nature of the feedback contribute to the overall improvement as improvement in translation expertise was streamlined through the scaffolding of each step and every session focused on a different area. Thus, as learners progressed through one session after another, they would gradually develop skills and these skills would again be integrated, piece by piece, by the end of the training program (68). Results in terms of improvements in ST comprehension indicate that Group A showed positive outcomes as to their awareness about a translator's responsibility for adequate ST comprehension since they could explain how they collected necessary information about ST and the strategies they employed (69). This skill connects to the one developed in the next session, that is information search skills. In session 3 the learners could only rely on a single online dictionary for finding equivalents, but in session 5 they could refer to multiple sources including authoritative information and Google search (69).

As to building a relationship with the client, learners in the cohort Group B indicated uncertainty about when and how to consult a client on the issues they deemed necessary in session 3; however through professional feedback, awareness- raising activities, and self-reflection, they could distinguish issues on which they needed to consult with a client from those requiring the translator's judgment and in session 5 the Professional reviewer indicated

that their sense of timing for asking for a translation brief and consulting the client was appropriate (70). The final key area in learner improvement was appropriate readability in a TT. Group B, who had less experience in a real-world scenario and therefore required a longer period of time to improve themselves, was also the group who struggled most in this area. They were uncertain how to consult a client for the purpose of the Project and presented the client with multiple equivalents of a term which caused the Professional reviewer to point out that they did not have adequate knowledge on what the responsibilities of a professional translator entailed (71). They were not aware either of the risk behind providing redundant additional information in the content, which signaled that they underestimated target readers' existing knowledge of a topic. When given these comments in Session 5, the learner group could gain a better understanding of how important it was to consult a client (71).

In the case of learners' perception of the usefulness of TBL, an outstanding benefit the learners commented on was about a better understanding of key market requirements. Beyond tasks that simply aim to find the correct TT equivalents, the model facilitated learning through making learners familiarize themselves with industry requirements that go beyond the mere translation of a text with "contextualized tasks reflecting a wide range of real-life challenges" (71-72). Another prominent benefit was about "the importance of understanding the concept of risk" while the next was building a good relationship with a client (72). To develop the related skills, learning outcomes depended on using experiential learning and needed to be "authentically situated", which points to practicum opportunities⁹⁶(72). In addition to the above benefits, the learners also suggested that the use of situated translation tasks and professional feedback enabled them to understand the relevance of theories to the professional environment (73). For instance, a learner remarked on "automatized knowledge and skills at various levels", which was consistent with the model of "Deliberate Practice" put forward by Ericsson (2000 cited in Inoue and Candlin 2015, 73). All comments underlined that having a learner-centered, collaborative

⁹⁶ Practicum opportunities are a part of the NAATI requirements in Australia.

learning/teaching environment helped them learn and raised their awareness of their weaknesses (74). Thus, the authors state that the assertion that model serves the purpose “is based on the evidence presented” and that “the approach adopted is pedagogically sound, authentic” and it “maximizes learner knowledge and skills in identifying and addressing key challenges by means of enhanced learner autonomy and peer collaboration” (74).

Using authentic learning material in a real-world scenario is a key principle the present study aims to achieve. As for the holistic approach to learning, we attempted to apply it through taking advantage of MT as an augmentor of learning to learn in the translation workflow. The priming effect that this study is designed to observe and measure is pivotal for exploring the nature of the collaborative environment in which today’s translators are expected to acquire skills. Thus, in the next part we present the results of the two-tiered priming experiment conducted for these purposes.

5.4.4 Post-editing evaluations

We applied mixed model analysis with ANOVA to the descriptive data belonging to 7 participants using IBM SPSS 2.0. The error annotations were carried out using Matecat’s own annotation tool with 5 issue types, i.e., Language quality (grammar, punctuation, spelling), Style (readability, consistent style and tone), Tag issues (mismatches, whitespaces), Terminology and Translation consistency and Translation errors (mistranslation, additions or omissions) in 3 severity levels – Major, Minor and Neutral. Therefore, in terms of the descriptive data we obtained from the experiment, we had 10 observations for each of the seven participants, making a total of 70 valid data sets for both the “Priming Set” and the “Test Set”. The data sets are evenly distributed, meaning that each consists of 50% (35 observations) of the “Priming Set” and “Test Set” data sets. This distribution ensures that the analysis is performed equally on both types of data. Table 7 below summarizes the total number of errors annotated for both sets under each issue type, with P signifying a Participant.

Table 6: Descriptive Statistics for All Participants

Total Score						
Participant	Data Set Type	Issues	Count	Mean	Standard Deviation	Coefficient of Variation
P1	Priming Set	Language quality (grammar, punctuation, spelling)	1	2,000	.	.
		Style (readability, consistent style and tone)	1	7,500	.	.
		Tag issues (mismatches, whitespaces)	1	,000	.	.
		Terminology and translation consistency	1	2,000	.	.
		Translation errors (mistranslation, additions or omissions)	1	2,500	.	.
		Total	5	2,800	2,7973	99,9%
	Test Set	Language quality (grammar, punctuation, spelling)	1	8,500	.	.
		Style (readability, consistent style and tone)	1	66,000	.	.
		Tag issues (mismatches, whitespaces)	1	6,000	.	.
		Terminology and translation consistency	1	2,000	.	.

		Translation errors (mistranslation, additions or omissions)	1	31,000	.	.
		Total	5	22,700	26,7058	117,6%
Total		Language quality (grammar, punctuation, spelling)	2	5,250	4,5962	87,5%
		Style (readability, consistent style and tone)	2	36,750	41,3657	112,6%
		Tag issues (mismatches, whitespaces)	2	3,000	4,2426	141,4%
		Terminology and translation consistency	2	2,000	,0000	0,0%
		Translation errors (mistranslation, additions or omissions)	2	16,750	20,1525	120,3%
		Total	10	12,750	20,7475	162,7%
P2	Priming Set	Language quality (grammar, punctuation, spelling)	1	4,000	.	.
		Style (readability, consistent style and tone)	1	2,000	.	.
		Tag issues (mismatches, whitespaces)	1	2,500	.	.

	Terminology and translation consistency	1	,000	.	.
	Translation errors (mistranslation, additions or omissions)	1	,500	.	.
	Total	5	1,800	1,6047	89,1%
Test Set	Language quality (grammar, punctuation, spelling)	1	3,000	.	.
	Style (readability, consistent style and tone)	1	9,500	.	.
	Tag issues (mismatches, whitespaces)	1	2,500	.	.
	Terminology and translation consistency	1	1,500	.	.
	Translation errors (mistranslation, additions or omissions)	1	17,500	.	.
	Total	5	6,800	6,7602	99,4%
Total	Language quality (grammar, punctuation, spelling)	2	3,500	,7071	20,2%
	Style (readability, consistent style and tone)	2	5,750	5,3033	92,2%

		Tag issues (mismatches, whitespaces)	2	2,500	,0000	0,0%
		Terminology and translation consistency	2	,750	1,0607	141,4%
		Translation errors (mistranslation, additions or omissions)	2	9,000	12,0208	133,6%
		Total	10	4,300	5,3292	123,9%
P3	Priming Set	Language quality (grammar, punctuation, spelling)	1	2,000	.	.
		Style (readability, consistent style and tone)	1	7,500	.	.
		Tag issues (mismatches, whitespaces)	1	,000	.	.
		Terminology and translation consistency	1	2,000	.	.
		Translation errors (mistranslation, additions or omissions)	1	2,500	.	.
		Total	5	2,800	2,7973	99,9%
	Test Set	Language quality (grammar, punctuation, spelling)	1	1,000	.	.

	Style (readability, consistent style and tone)	1	,000	.	.
	Tag issues (mismatches, whitespaces)	1	2,000	.	.
	Terminology and translation consistency	1	,000	.	.
	Translation errors (mistranslation, additions or omissions)	1	33,500	.	.
	Total	5	7,300	14,6697	201,0%
Total	Language quality (grammar, punctuation, spelling)	2	1,500	,7071	47,1%
	Style (readability, consistent style and tone)	2	3,750	5,3033	141,4%
	Tag issues (mismatches, whitespaces)	2	1,000	1,4142	141,4%
	Terminology and translation consistency	2	1,000	1,4142	141,4%
	Translation errors (mistranslation, additions or omissions)	2	18,000	21,9203	121,8%
	Total	10	5,050	10,2346	202,7%

P4	Priming Set	Language quality (grammar, punctuation, spelling)	1	2,500	.	.
		Style (readability, consistent style and tone)	1	10,500	.	.
		Tag issues (mismatches, whitespaces)	1	1,000	.	.
		Terminology and translation consistency	1	,500	.	.
		Translation errors (mistranslation, additions or omissions)	1	2,000	.	.
		Total	5	3,300	4,1018	124,3%
	Test Set	Language quality (grammar, punctuation, spelling)	1	22,500	.	.
		Style (readability, consistent style and tone)	1	6,500	.	.
		Tag issues (mismatches, whitespaces)	1	35,000	.	.
		Terminology and translation consistency	1	2,000	.	.
		Translation errors (mistranslation, additions or omissions)	1	28,000	.	.

		Total	5	18,800	14,0917	75,0%
	Total	Language quality (grammar, punctuation, spelling)	2	12,500	14,1421	113,1%
		Style (readability, consistent style and tone)	2	8,500	2,8284	33,3%
		Tag issues (mismatches, whitespaces)	2	18,000	24,0416	133,6%
		Terminology and translation consistency	2	1,250	1,0607	84,9%
		Translation errors (mistranslation, additions or omissions)	2	15,000	18,3848	122,6%
		Total	10	11,050	12,7464	115,4%
P5	Priming Set	Language quality (grammar, punctuation, spelling)	1	3,000	.	.
		Style (readability, consistent style and tone)	1	12,500	.	.
		Tag issues (mismatches, whitespaces)	1	,500	.	.
		Terminology and translation consistency	1	4,500	.	.

	Translation errors (mistranslation, additions or omissions)	1	,500	.	.
	Total	5	4,200	4,9447	117,7%
Test Set	Language quality (grammar, punctuation, spelling)	1	16,500	.	.
	Style (readability, consistent style and tone)	1	19,500	.	.
	Tag issues (mismatches, whitespaces)	1	,500	.	.
	Terminology and translation consistency	1	13,500	.	.
	Translation errors (mistranslation, additions or omissions)	1	16,000	.	.
	Total	5	13,200	7,4128	56,2%
Total	Language quality (grammar, punctuation, spelling)	2	9,750	9,5459	97,9%
	Style (readability, consistent style and tone)	2	16,000	4,9497	30,9%
	Tag issues (mismatches, whitespaces)	2	,500	,0000	0,0%

		Terminology and translation consistency	2	9,000	6,3640	70,7%
		Translation errors (mistranslation, additions or omissions)	2	8,250	10,9602	132,9%
		Total	10	8,700	7,6019	87,4%
P6	Priming Set	Language quality (grammar, punctuation, spelling)	1	2,000	.	.
		Style (readability, consistent style and tone)	1	2,000	.	.
		Tag issues (mismatches, whitespaces)	1	2,000	.	.
		Terminology and translation consistency	1	2,000	.	.
		Translation errors (mistranslation, additions or omissions)	1	4,000	.	.
		Total	5	2,400	,8944	37,3%
	Test Set	Language quality (grammar, punctuation, spelling)	1	25,500	.	.
		Style (readability, consistent style and tone)	1	,000	.	.

		Tag issues (mismatches, whitespaces)	1	,000	.	.
		Terminology and translation consistency	1	2,000	.	.
		Translation errors (mistranslation, additions or omissions)	1	31,500	.	.
		Total	5	11,800	15,4135	130,6%
Total		Language quality (grammar, punctuation, spelling)	2	13,750	16,6170	120,9%
		Style (readability, consistent style and tone)	2	1,000	1,4142	141,4%
		Tag issues (mismatches, whitespaces)	2	1,000	1,4142	141,4%
		Terminology and translation consistency	2	2,000	,0000	0,0%
		Translation errors (mistranslation, additions or omissions)	2	17,750	19,4454	109,6%
		Total	10	7,100	11,4232	160,9%
P7	Priming Set	Language quality (grammar, punctuation, spelling)	1	1,000	.	.

	Style (readability, consistent style and tone)	1	5,000	.	.
	Tag issues (mismatches, whitespaces)	1	6,500	.	.
	Terminology and translation consistency	1	,500	.	.
	Translation errors (mistranslation, additions or omissions)	1	11,500	.	.
	Total	5	4,900	4,4917	91,7%
Test Set	Language quality (grammar, punctuation, spelling)	1	,000	.	.
	Style (readability, consistent style and tone)	1	4,500	.	.
	Tag issues (mismatches, whitespaces)	1	1,500	.	.
	Terminology and translation consistency	1	,000	.	.
	Translation errors (mistranslation, additions or omissions)	1	,000	.	.
	Total	5	1,200	1,9558	163,0%

Total		Language quality (grammar, punctuation, spelling)	2	,500	,7071	141,4%
		Style (readability, consistent style and tone)	2	4,750	,3536	7,4%
		Tag issues (mismatches, whitespaces)	2	4,000	3,5355	88,4%
		Terminology and translation consistency	2	,250	,3536	141,4%
		Translation errors (mistranslation, additions or omissions)	2	5,750	8,1317	141,4%
		Total	10	3,050	3,8039	124,7%
Total	Priming Set	Language quality (grammar, punctuation, spelling)	7	2,357	,9449	40,1%
		Style (readability, consistent style and tone)	7	6,714	4,0089	59,7%
		Tag issues (mismatches, whitespaces)	7	1,786	2,2887	128,2%
		Terminology and translation consistency	7	1,643	1,5197	92,5%
		Translation errors (mistranslation, additions or omissions)	7	3,357	3,7938	113,0%

	Total	35	3,171	3,2357	102,0%
Test Set	Language quality (grammar, punctuation, spelling)	7	11,000	10,5198	95,6%
	Style (readability, consistent style and tone)	7	15,143	23,3946	154,5%
	Tag issues (mismatches, whitespaces)	7	6,786	12,5925	185,6%
	Terminology and translation consistency	7	3,000	4,7170	157,2%
	Translation errors (mistranslation, additions or omissions)	7	22,500	12,0968	53,8%
	Total	35	11,686	14,8646	127,2%
Total	Language quality (grammar, punctuation, spelling)	14	6,679	8,4617	126,7%
	Style (readability, consistent style and tone)	14	10,929	16,7077	152,9%
	Tag issues (mismatches, whitespaces)	14	4,286	9,0739	211,7%
	Terminology and translation consistency	14	2,321	3,4396	148,2%

Translation errors (mistranslation, additions or omissions)	14	12,929	13,1469	101,7%
Total	70	7,429	11,5075	154,9%

Starting with the data from quality scores, one of the foremost results we noticed about the quality scores is that there is a high variation across participants. However, the overall quality score is generally significantly higher (or more consistent) in the test set than in the priming set. Issue or error categories in the test set tend to differ across participants and it is clear that some participants are more inclined to annotate errors in certain issue types in both the priming and the test set. When it comes to assessing the results from test set in terms of quality, the highest number of issues was annotated in the categories of Style and Translation Errors in the test set, the number of errors annotated by all participants was dramatically higher compared to the priming set and the overall quality was much lower across all participants. Only one participant found the Kantan MT output to be of higher quality after the priming set. Surprisingly, the number of errors annotated under the category Terminology and translation consistency was lower in the test set. Yet, the distribution of errors annotated under Terminology is more consistent in the priming set compared to test set where one participant had 13.5 error points in total for this single error category. The highest number of errors annotated for the test set was found in the issue types Style and Translation Errors. When compared, the test set is significantly different from the priming set in terms of quality scores. In particular, the high variances of P5 and P6 in the categories Style and Translation Errors indicate that more errors are annotated in the test set, resulting in a lower total quality score. Table 2 demonstrates that the issue type Translation Errors has lower variance and consistency, whereas Style and Tag errors are distributed with higher variance in both the priming and the test set, suggesting that participants experience more inconsistency in their tagging. It must be noted here that because the MT output does not differentiate between tagged and non-tagged items in a text and works on a plain UTF-8

format, copying the output in the project file, we used Matecat’s “Guess Tag” option that the tool suggests if it detects tags in the ST before hitting the “Translated” button. Therefore, as tags in both Tasks were automatically inserted into the segments, the results for the high number of errors annotated under Style Tag errors were not surprising but we did not expect the variance to be high. This result might be interpreted as the outcome of participants’ intuition that style is not going to be considered as a crucial error category to be kept in mind because the ‘linguistic’ content is probably going to undergo a graphic design process after completion.

On the other hand, Language quality and Style, i.e., consistency, are the categories with the highest variance in both priming and test sets, indicating significant differences in terms of error annotations among participants in both sets. These data suggests that there are large differences in the quality scores across participants in general, with inconsistencies (or rather differences) being more pronounced for some issue types. Figure 8 below demonstrate the error issue types and the number of errors annotated under them by each participant.

Table 7: Error Issue Type Annotations

Participant 1									
Issues (Priming Set)	Major (Weight:2)	Minor (Weight: 0.5)	Neutral (Weight: 0)	Total	Issues (Test Set)	Major (W:2)	Minor (W:0.5)	Neutral (W:0)	Total
Style (readability, consistent style and tone)	0	15		3 7.5	Style (readability, consistent style and tone)	33	0		1 66
Tag issues (mismatches, whitespaces)	0	0		2 0	Tag issues (mismatches, whitespaces)	3	0		5 6
Translation errors (mistranslation, additions or omission)	1	1		1 2.5	Translation errors (mistranslation, additions or omission)	14	6		7 31
Terminology and translation consistency	0	4		0 2	Terminology and translation consistency	1	0		2 2
Language quality (grammar, punctuation, spelling)	0	4		6 2	Language quality (grammar, punctuation, spelling)	4	1		7 8.5
Total Error Points				14	Total Error Points				113.5
Quality Score	23.26				Quality Score	188.54			
Total Time Spent	00:11				Total Time Spent	01:46:26			
Post-Editing Effort	96				Post-Editing Effort	98			
Participant 2									
Issues (Priming Set)	Major	Minor	Neutral	Total	Issues (Test Set)	Major	Minor	Neutral	Total
Style (readability, consistent style and tone)	0	4		2 2	Style (readability, consistent style and tone)	3	7		2 9.5
Tag issues (mismatches, whitespaces)	0	1		1 2.5	Tag issues (mismatches, whitespaces)	1	1		6 2.5
Translation errors (mistranslation, additions or omission)	0	1		1 0.5	Translation errors (mistranslation, additions or omission)	6	11		3 17.5
Terminology and translation consistency	0	0		0	Terminology and translation consistency	0	3		2 1.5
Language quality (grammar, punctuation, spelling)	1	4		4 4	Language quality (grammar, punctuation, spelling)	0	6		2 3
Total Error Points				9	Total Error Points				34
Quality Score	14.95				Quality Score	58.02			
Total Time Spent	00:00				Total Time Spent	00:41			
Post-Editing Effort	99				Post-Editing Effort	00:00:00			

Participant 3									
Issues (Priming Set)					Issues (Test Set)				
Style (readability, consistent style and tone)	Major	Minor	Neutral	Total	Style (readability, consistent style and tone)	Major	Minor	Neutral	Total
Tag issues (mismatches, whitespaces)		0	15	3 7.5	Style (readability, consistent style and tone)		0	0	0 0
Translation errors (mistranslation, additions or omission)		0	0	2 0	Tag issues (mismatches, whitespaces)		1	0	0 2
Terminology and translation consistency		1	1	1 2.5	Translation errors (mistranslation, additions or omission)		15	7	1 33.5
Language quality (grammar, punctuation, spelling)		0	4	0 2	Terminology and translation consistency		0	0	0 0
Total Error Points		0	4	6 2	Language quality (grammar, punctuation, spelling)		0	2	0 1
Quality Score	23.26			14	Total Error Points				60.63
Total Time Spent	00:00:08				Quality Score	60.63			
Post-Editing Effort	96				Total Time Spent	00:32:23			
					Post-Editing Effort	98			
Participant 4									
Issues (Priming Set)					Issues (Test Set)				
Style (readability, consistent style and tone)	Major	Minor	Neutral	Total	Style (readability, consistent style and tone)	Major	Minor	Neutral	Total
Tag issues (mismatches, whitespaces)		3	9	0 10.5	Style (readability, consistent style and tone)		3	1	0 6.5
Translation errors (mistranslation, additions or omission)		0	2	0 1	Tag issues (mismatches, whitespaces)		17	2	0 55
Terminology and translation consistency		0	4	0 2	Translation errors (mistranslation, additions or omission)		13	4	0 28
Language quality (grammar, punctuation, spelling)		0	1	0 0.5	Terminology and translation consistency		1	0	0 2
Total Error Points		0	5	0 2.5	Language quality (grammar, punctuation, spelling)		11	1	0 22.5
Quality Score	27.41			16.5	Total Error Points				94
Total Time Spent	02:38:04				Quality Score	156.15			
Post-Editing Effort	99				Total Time Spent	01:45:09			
					Post-Editing Effort	97			
Participant 5									
Issues (Priming Set)					Issues (Test Set)				
Style (readability, consistent style and tone)	Major	Minor	Neutral	Total	Style (readability, consistent style and tone)	Major	Minor	Neutral	Total
Tag issues (mismatches, whitespaces)		2	17	10 12.5	Style (readability, consistent style and tone)		5	19	10 19.5
Translation errors (mistranslation, additions or omission)		0	1	0 0.5	Tag issues (mismatches, whitespaces)		0	1	0 0.5
Terminology and translation consistency		0	1	0 0.5	Translation errors (mistranslation, additions or omission)		5	12	1 56
Language quality (grammar, punctuation, spelling)		1	5	0 4.5	Terminology and translation consistency		6	3	1 13.5
Total Error Points		0	6	1 3	Language quality (grammar, punctuation, spelling)		3	21	0 16.5
Quality Score	34.88			21	Total Error Points				66
Total Time Spent	03:55:48				Quality Score	109.82			
Post-Editing Effort	0				Total Time Spent	15:05:09			
					Post-Editing Effort	100			
Participant 6									
Issues (Priming Set)					Issues (Test Set)				
Style (readability, consistent style and tone)	Major	Minor	Neutral	Total	Style (readability, consistent style and tone)	Major	Minor	Neutral	Total
Tag issues (mismatches, whitespaces)		0	4	23 2	Style (readability, consistent style and tone)		0	0	7 0
Translation errors (mistranslation, additions or omission)		0	4	0 2	Tag issues (mismatches, whitespaces)		0	0	0 0
Terminology and translation consistency		0	8	0 4	Translation errors (mistranslation, additions or omission)		14	7	0 31.5
Language quality (grammar, punctuation, spelling)		0	4	2 2	Terminology and translation consistency		1	0	2 2
Total Error Points		0	4	0 2	Language quality (grammar, punctuation, spelling)		12	3	1 25.5
Quality Score	18.24			11	Total Error Points				59
Total Time Spent	00:01:25				Quality Score	98.01			
Post-Editing Effort	45				Total Time Spent	00:24			
					Post-Editing Effort	53			
Participant 7									
Issues (Priming Set)					Issues (Test Set)				
Style (readability, consistent style and tone)	Major	Minor	Neutral	Total	Style (readability, consistent style and tone)	Major	Minor	Neutral	Total
Tag issues (mismatches, whitespaces)		1	6	0 5	Style (readability, consistent style and tone)		0	9	10 4.5
Translation errors (mistranslation, additions or omission)		3	1	0 6.5	Tag issues (mismatches, whitespaces)		0	3	0 1.5
Terminology and translation consistency		5	3	0 11.5	Translation errors (mistranslation, additions or omission)		0	0	0 0
Language quality (grammar, punctuation, spelling)		0	1	0 0.5	Terminology and translation consistency		0	0	0 0
Total Error Points		0	2	0 1	Language quality (grammar, punctuation, spelling)		0	0	2 0
Quality Score	41.11			24.5	Total Error Points				6
Total Time Spent	00:55:13				Quality Score	17			
Post-Editing Effort	54				Total Time Spent	00:00:09			
					Post-Editing Effort	91			

In order to understand the impact of error issue types on total number of errors for both sets, Table 8 below lays out the results obtained using a fixed effects analysis in a linear regression model with quality scores for both sets as the dependent variable.

Table 8: Test of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	18,897	36,935	,000
DataSetType	1	12,928	6,690	,023
Issues	4	9,854	3,390	,054

a. Dependent Variable: Total Score.

When we measure the intercept⁹⁷, the F-value obtained is very high ($f=36.935$) and also very significant with a p-value of 0.000. This indicates that the fixed effect of the model is statistically significant and that the independent variables serve as a starting point. In other words, all effects in the model indicate that the intercept is significant and meaningful for interpreting the data. In the case of DataSetType, the F-value obtained for the DataSetType variable is 6.690 and the p-value is 0.023, showing that there is a significant relationship between the data set type (e.g. test set vs. priming set) and Total Error Points for each issue type. That is, the dataset type has a significant effect on the total score. This effect is significant at $p < 0.05$ level.

In terms of issues, or error types, the F-value obtained for the effect of the Issues variable is 3.390 and the p-value is 0.054. This p-value is very close to the 0.05 significance limit, but still does not exceed the 0.05 threshold. This demonstrates that error types can have an effect on the total score for all errors, but this effect is statistically borderline and a clearer result can be reached with a larger sample. Overall, the Intercept and DataSetType independent variables have a significant effect on the total score ($p < 0.05$), while the Issues variable is at the significance limit with a p-value of 0.054, indicating that error types are likely to have an impact on the total score of errors, but further research is needed to determine a definitive effect. For this reason, we also measured estimates of fixed effects where DataSetType Test set was taken as the reference and hence, was given the value 0 as demonstrated in Table 10 below. Here the priming set score is -4,28 points lower than the test set's average score and the P value of 0,023 is statistically meaningful. For instance, the issue type Language quality was estimated to be -1,59, meaning Language quality has a negative impact on the total score for errors but because p is ,312, it is not statistically significant. Interpreting the results obtained from fixed effects estimates with the issue types and categories for each participant reveals that the second hypothesis, which states *the*

⁹⁷ Bobbitt explains that intercept which is also called a 'constant' "represents the mean value of the response variable when all of the predictor variables in the model are equal to zero" in a regression model" (Bobbitt 2021).

number of errors annotated in the test set will be less in the issue types annotated in the priming set, is not confirmed and on the contrary, the numbers of errors annotated in the test set were much higher and considered to be Major in more instances compared to Minor or Neutral severity levels most errors were annotated in Task1.

Table 9: Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	8,270451	2,128661	20,566	3,885	,001	3,837963	12,702939
[DataSetType=Priming Set]	-4,281543	1,655397	12,928	-2,586	,023	-7,859834	-,703252
[DataSetType=Test Set]	0 ^b	0	.	.	.	.	.
[Issues=Language quality (grammar, punctuation, spelling)]	-1,598558	1,465899	6,977	-1,090	,312	-5,067150	1,870034
[Issues=Style (readability, consistent style and tone)]	2,856582	2,057678	12,659	1,388	,189	-1,600948	7,314111
[Issues=Tag issues (mismatches, whitespaces)]	-2,177348	1,654317	10,207	-1,316	,217	-5,853267	1,498571
[Issues=Terminology and translation consistency]	-2,565308	1,527376	8,191	-1,680	,131	-6,073215	,942599
[Issues=Translation errors (mistranslation, additions or omissions)]	0 ^b	0	.	.	.	.	.
a. Dependent Variable: Total Score.							

5.4.5 Statistics for the difference between the priming set and the test set

The independent samples t-test analysis examines the difference in total error points between priming and test sets. Levene's test of homogeneity of variance indicates rejection of the assumption of equal variances for both groups ($p = 0.000$), indicating that there are different

variance levels between the groups. However, the t-test results reveal that the mean difference between the two groups (Test Set: 11.686 - Priming Set: 3.171 = -8.5143) is statistically significant ($p = 0.001$). This indicates that the priming set has a significantly lower total score of errors than the test set and that this difference reflects the true difference between the groups. These findings emphasize that the differences in the performances of the priming and test sets are significant. These results refute the second hypothesis that states *the number of total errors will be higher in the test set than the priming set*.

5.4.6 Time spent on post-editing

Table 11 below presents the descriptive statistics of various variables (Total Score for Errors, Quality Score, Total Time Spent and Post-Editing Effort) for two different datasets (Priming Set and Test Set). In terms of the total score for errors, the mean is 3.171, the standard deviation is 3.2357, and the 95% confidence interval varies from 2.060 to 4.283 for the priming set. On the other hand, for the test set, the mean is 11.686, the standard deviation is 14.8646, and the 95% confidence interval varies from 6.580 to 16.792, indicating that the test set has a higher mean and the total score for errors presents a large variance. The difference between the two sets is quite significant and a much wider distribution is observed in the Test Set.

In terms of the quality scores, the mean is 26.1586, the standard deviation is 8.6289, the 95% confidence interval varies between 23.1944 and 29.1227 in the priming set. In test set, the mean is 98.3100, the standard deviation is 55.9330, and the 95% confidence interval varies between 17.5237 and 117.5237. The mean value of the test set is quite high compared to the Priming Set and the variance is also significantly higher, indicating that the Test Set received much more diverse evaluations.

Table 10: Descriptive Statistics of Total Quality, Time Spent on PE and PE Effort

Descriptives ^a										
		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
Total Score	Priming Set	35	3,171	3,2357	,5469	2,060	4,283	,0	12,5	
	Test Set	35	11,686	14,8646	2,5126	6,580	16,792	,0	66,0	
	Total	70	7,429	11,5075	1,3754	4,685	10,172	,0	66,0	
	Model Fixed Effects			10,7570	1,2857	4,863	9,994			
	Random Effects				4,2571	-46,664	61,521			32,9404
T. Total Score	Priming Set	35	15,714	5,1296	,8671	13,952	17,476	9,0	24,5	
	Test Set	35	58,429	34,6881	5,8634	46,513	70,344	6,0	113,5	
	Total	70	37,071	32,6897	3,9072	29,277	44,866	6,0	113,5	
	Model Fixed Effects			24,7950	2,9636	31,158	42,985			
	Random Effects				21,3571	-234,297	308,440			894,6897
Quality Score	Priming Set	35	26,1586	8,62890	1,45855	23,1944	29,1227	14,95	41,11	
	Test Set	35	98,3100	55,93309	9,45442	79,0963	117,5237	17,00	188,54	
	Total	70	62,2343	53,83858	6,43494	49,3969	75,0716	14,95	188,54	
	Model Fixed Effects			40,01855	4,78313	52,6897	71,7789			
	Random Effects				36,07571	-396,1511	520,6197			2557,15763
Total Time Spent	Priming Set	35	1:05:57	1:28:32	0:14:57	0:35:32	1:36:22	0:00:05	3:55:48	
	Test Set	35	2:53:28	5:05:25	0:51:37	1:08:32	4:38:23	0:00:09	15:05:09	
	Total	70	1:59:42	3:49:41	0:27:27	1:04:56	2:54:29	0:00:05	15:05:09	
	Model Fixed Effects			3:44:51	0:26:52	1:06:05	2:53:20			
	Random Effects				0:53:45	-9:23:16	13:22:42			4333:59
Post-Editing Effort	Priming Set	35	69,86	36,004	6,086	57,49	82,23	0	99	
	Test Set	35	76,71	35,379	5,980	64,56	88,87	0	100	
	Total	70	73,29	35,601	4,255	64,80	81,77	0	100	
	Model Fixed Effects			35,693	4,266	64,77	81,80			
	Random Effects				4,266 ^a	19,08 ^a	127,49 ^a			-12,890

a. Warning: Between-component variance is negative. It was replaced by 0.0 in computing these random effects measure.

In case of the total time spent on PE, the mean is 1:05:57 (1 hour 5 minutes 57 seconds), the standard deviation is 1:28:32, and the 95% confidence interval varies between 0:35:32 and 1:36:22 in the priming set, while in the test Set the mean is 2:53:28 (2 hours 53 minutes 28 seconds), the standard deviation is 5:05:25, and the 95% confidence interval ranges from 2:54:29 to 3:55:48. The time spent on the PE process of the test set is significantly longer and the variance measured is also higher, indicating that the test set required more time and that participants worked at different speeds.

In the case of the post-editing effort, the mean is 69.86, the standard deviation is 36.004, and the 95% confidence interval ranges from 57.49 to 82.23 in the priming set, whereas the mean is 76.71, the standard deviation is 35.379, and the 95% confidence interval ranges from 64.56 to 88.87 in the test set. The editing effort on the test set is greater and presents a wider range. Overall, the test set generally presents higher means and more variance, indicating that participants put in more effort and spent more time on it. Additionally, there are significant differences between quality score and total time spent

between the test set and the priming set. Analysis of these differences is critical to understanding the validity of the experiment and how performance changes under different conditions.

One final test before ANOVA is a paired samples t-test where the relationship between the total score for errors, the PE effort and the time spent on PE is analyzed as seen in Table 11 below:

Table 11: Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Total Score - Post-Editing Effort	-65,8571	36,2111	4,3280	-74,4914	-57,2229	-15,216	69	,000
Pair 2	Post-Editing Effort - Total Time Spent	-7109,571	13777,967	1646,782	-10394,812	-3824,331	-4,317	69	,000
Pair 3	Total Score - Quality Score	-25,16286	21,19327	2,53308	-30,21621	-20,10950	-9,934	69	,000

The first test analyzing the difference between the total score and the post-editing effort shows that the difference between these two variables is significant ($p = 0.000$), with a mean difference of -65.8571 and a confidence interval ranging from -74.4914 to -57.2229. These results indicate that the post-editing effort is significantly different from the total score of errors. The second test analyzes the difference between the post-editing effort and the time spent. Here too, a significant difference was found ($p = 0.000$), with a mean difference of -

7109.571 and a confidence interval ranging from -10394.812 to -3824.331. These results clearly demonstrate that the time spent on PE is significantly different from that spent on the post-editing effort. The third test examining the relationship between the total score and the quality score also yielded a very strong correlation ($r = 0.999$, $p = 0.000$), indicating an almost perfect relationship between the total score for errors and the quality score. The mean difference is -25.16268, the confidence interval is between -30.21621 and -20.10950, the difference is significant ($p = 0.000$). In conclusion, these analyses reveal that there is a strong relationship between the total score for errors and the quality score, and there are significant differences between the post-editing effort and time spent on PE.

Finally, Table 13 below shows the results of the ANOVA (Analysis of Variance) test which “compares the means of three or more independent groups to determine if there is a statistically significant difference between the corresponding population means” (Bobbitt 2018). The table presents Between Groups and Within Groups analysis for four different variables, with groups referring to priming and test sets.

Table 12: ANOVA Test Results

		Sum of Squares	df	Mean Square	F	Sig.
Total Score	Between Groups	1268,629	1	1268,629	10,964	,001
	Within Groups	7868,514	68	115,713		
	Total	9137,143	69			
T.Total Score	Between Groups	31928,929	1	31928,929	51,935	,000
	Within Groups	41805,714	68	614,790		
	Total	73734,643	69			
Quality Score	Between Groups	91102,001	1	91102,001	56,886	,000
	Within Groups	108900,925	68	1601,484		
	Total	200002,927	69			
Total Time Spent	Between Groups	728108251,429	1	728108251,429	4,000	,049
	Within Groups	12377768887,143	68	182026013,046		
	Total	13105877138,571	69			
Post-Editing Effort	Between Groups	822,857	1	822,857	,646	,424
	Within Groups	86631,429	68	1273,992		

	Total	87454,286	69			
--	-------	-----------	----	--	--	--

In terms of total score for errors, the F value is 10.964 and the p value is 0.001, indicating a significant difference between the groups ($p < 0.05$). In other words, in terms of total score, there is a statistically significant difference between the data set types (Priming Set and Test Set). In the case of total time spent, the F value is 4.000 and the p value is 0.049, again expressing a statistically significant difference ($p < 0.05$). The results can also be interpreted that in terms of total time spent on PE, where there is a significant difference between the groups. Yet, this difference may have a lower effect than the total score and quality score. In case of quality score, the F value is 56.886 and the p value is 0.000, showing a very strongly significant difference between the groups. Therefore, in terms of quality score, there is a large difference between the groups ($p < 0.05$), indicating that the dataset types make a significant difference in quality assessment. Lastly, for the data obtained from post-editing effort the F value is 0.646 and the p value is 0.424. These results indicate that there is no significant difference between the groups ($p > 0.05$) in terms of post-editing effort. Overall, though significant differences between the groups were observed in terms of total score for errors and quality score, a significant difference was also observed in terms of total time spent. Interestingly, no significant difference was observed between the priming set and test set in terms of PE effort. After these statistical differences, we can calculate Cohen's d^{98} using the data we obtained as a result of ANOVA, where means of both groups (M1 (Priming Set) and M2 (Test Set)), standard deviations of both groups (SD1, SD2) and sample sizes of both groups (n_1, n_2) are calculated with the definition Cohen's $d = (M1 - M2) / SD_{pooled}$ as demonstrated in Table 13 below:

Table 13: Calculations for Cohen's D

M1 (Priming Set): Mean (M1) = 3.171	M2 (Test Set): Mean (M2) = 11.686
--	--------------------------------------

⁹⁸ Piasta and Justice explain Cohen's d as follows: "As an effect size, Cohen's d is typically used to represent the magnitude of differences between two (or more) groups on a given variable, with larger values representing a greater differentiation between the two groups on that variable" (2010).

Standard Deviation (SD1) = 3.2357 Sample Size (n1) = 35	Standard Deviation (SD2) = 14.8646 Sample Size (n2) = 35
--	---

The result of the calculation is as below:

$$\text{Cohen's } d \approx -0.832$$

This result shows a large effect because the absolute value of Cohen's d is above 0.8, pointing to a large difference between the two groups, meaning that the mean of the test set is significantly higher than the priming set. This result also confirms the first half of the fourth hypothesis which states *translators will spend more time on post-editing in Task 2*.

5.4.7 The comparison of novices and professionals

In order to compare the results of the priming effect on novices and the professionals separately, the data obtained from time spent on PE were analyzed, followed by the translations of the specific priming sentences. The first hypothesis of the present study, suggesting that *the post-editing by less experienced translators will be faster compared to professional translators because professionals would be attuned to errors and would need a longer period of post-editing time*, was not confirmed, as Tables 14 for the cohort of experienced professionals (P 3,6 and 7) and Table 15 for the cohort of less experienced groups demonstrate:

Table 14: Time Spent on PE by Professionals

	Participant 3		Participant 6		Participant 7	
	Priming Set	Test Set	Priming Set	Test Set	Priming Set	Test Set
Total Time Spent	00:00:08	00:32:23	01:25	00:24	00:55:13	00:00:09

Table 15: Time Spent on PE by the Less Experienced Group

	Participant 1		Participant 2		Participant 4		Participant 5	
	Priming Set	Test Set	Priming Set	Test Set	Priming Set	Test Set	Priming Set	Test Set
Total Time Spent	00:11	01:46:26	00:00	00:41	02:38:04	01:45:09	03:55:48	15:05:09

Therefore, for comparison on the textual level, the questions below were asked:

- Did professionals preserve the priming expressions when post-editing the priming set and the test set?
- Did the less-experienced group preserve the priming expressions when post-editing the priming set and the test set?

Participant 3 post-edited 11 segments out of 24 primers and her revisions were mostly evaluated under the issue Language Quality (grammar, punctuation and spelling) in the priming set. In the test set, there is a very clear pattern that expresses she worked on the revisions of the previously primed sentences. But in the test set, she evaluated most errors under the issue Translation Error in the major category, hence the quality score for the test set was drastically lower than for the priming set. Another quite experienced participant, the most experienced among the professionals' group with more than 7+ years of expertise in the industry, revised almost 20 segments out of all the primers including those we deliberately rewrote with a passive structure, though the ST clearly presented an active verb formation. He revised some of the primers which have perfect tense markers at the end of the main verb in the sentence with the non-primed progressive tense version. He also evaluated segments under more than one single issue type. For the test set, it can be argued that he was careful about revising the sentences with unnecessary passive structures. The final participant in the experienced group did not revise the primed segment in a very detailed way; however, she changed some verbs in MT output with more commonly accepted semantic collocations. She again revised the progressive tense marker at the end of some verbs with the aorist, i.e., broad tense marker, for present tense, and general events, and preferred the active subject marker

for some sentences primed with a passive voice. Also, similarly to the previous participant, she evaluated some segments with more than one issue type. In the test set, her evaluation resulted in a much higher quality score with a lower score for errors, marking the MT output with a Pass quite curiously.

In the novice group, on the other hand, Participant 1 revised the passive primers with an active voice structure and also opted for the postposition “için”, instead of the conjunction “amaçlı” in the first segment. Instead of using the aorist throughout the entire document, she preferred using the more literal counterpart of progressive tense. Moreover, it is clear that she chose the active voice if she deemed the ST would permit it, while revising most of the primed segments. However, compared to the more experienced cohort, apparently, she was reticent in her evaluations in the priming set. As for the test set, she revised all the segments apart from a subheading, always with an active voice structure. She also evaluated segments with more than one issue type if she thought it fitting. This participant provided additional feedback for Task 2 where she stated “this translation is unacceptable in that the characters are unreadable in Turkish, the translation is full of critical errors, the terminology is not consistent and the grammar rules are overlooked in the target language”. Below is an example from her revision for post-editing the test set in Task 2:

Figure 8: A Revised Segment by Participant 2

Source	We assess the importance and impact of the topics for the company, environment and society, and also for our stakeholders.	Words:	20
Suggestion			
Translation	Şirket, ortam ve toplum konularının önemli ve etkisini ve ayrıca paydaşlarımız için önemli ölçüde değerlendirilmektedir.	TTE:	09"
Revision	Şirketin, ortam, çevre ve toplum konularının önemli ve etkisini ve ayrıca paydaşlarımız için önemli ölçüde ve etkisini değerlendirilmektedir. yoruz.	TTE:	02'20"
QA	Human (3) Style (readability, consistent style and tone): [Major] Language quality (grammar, punctuation, spelling): [Major] Translation errors (mistranslation, additions or omissions): [Major]		

As for Participant 4, though her revision for the priming set in Task 1 was limited, she was keen to revise the primed segments in the passive voice with an active voice as the Figure 9 below shows.

Figure 9: A Revised Segment by Participant 4

4559807002	Secs/Word: 00" PEE: 0%	Segment status APPROVED
Source	We assess the importance and impact of the topics for the company, environment and society, and also for our stakeholders.	Words: 20
Suggestion		
Translation	Şirket, ortam ve toplum konularının önemli ve etkisini ve ayrıca paydaşlarımız için önemli ölçüde değerlendirilmektedir.	TTE: 10"
Revision	Şirket, ortam ve toplum konularının önemli ve etkisini ve ayrıca paydaşlarımız için önemli ölçüde değerlendirilmektedir.	TTE: 02"
QA	Human (1) Translation errors (mistranslation, additions or omissions): [Major]	

The final participant revised the primed segments in Task 1 by changing the conjunction as seen below:

Figure 10: A Revised Segment by Participant 5

4559264199	Secs/Word: 03" PEE: 0%	Segment status APPROVED
Source	A strong framework for sustainability.	Words: 5
Suggestion		
Translation	Sürdürülebilirlik amaçlı güçlü bir çerçeve.	TTE: 18"
Revision	Sürdürülebilirlik amaçlı güçlü ilişkin başarılı bir çerçeve.	TTE: 04'42"
QA	Human (1) Style (readability, consistent style and tone): [Major]	

He conducted meticulous research for the subject matter and changed the abbreviation of the concept “ESG” which stands for “environmental, social and governance” to “ÇSY”⁹⁹, the Turkish equivalent of the term. He was the only person in the experiment who searched for the term and provided a Turkish equivalent. He also revised the primed sentences in a passive voice with an active one as seen below in Figure 11.

⁹⁹ “ÇSY” is the abbreviation of “çevresel, sosyal ve kurumsal yönetim”.

Figure 11: A Revised Segment by Participant 5

Source	By addressing the three aspects of ESG, we are building a better future that helps us to	Words:	17
Suggestion			
Translation	ESG'nin üç yönünü ele alarak bize şu konularda yardımcı olan daha iyi bir gelecek inşa edilecektir.	TTE:	26"
Revision	ESG'nin üç yönünü ele alarak bize yaşadığımız konularda yardımcı olan hac daha iyi bir gelecek inşa edilecektir.	TTE:	02'40"
QA	Human (1)	Terminology and translation consistency: [Minor]	

Yet, overall the revisions suggest that he tried to keep his alterations to a minimum and for some segments he had more than one issue annotated. His revision of the same segment below in the test set provides an intriguing case, as we can observe the errors he considers in different severity levels, with examples:

Figure 12: The Same Segment Annotated in the Test Set

4572636444	Secs/Word: 00" PEE: 0%	Segment status APPROVED
Source	By · addressing · the · three · aspects · of · ESG, · we · are · building · a · better · future · that · helps · us · to ·	Words: 17
Suggestion		
Translation	 ESG'nin · üç · unsurunu · ele · alarak · bize · yardımcı · olmaya · yardımcı · olan · daha · iyi · bir · gelecek · oluşturunuz ·	TTE: 09"
Revision	 ESG'nin · üç · unsurunu · ele · alarak · bize · yardımcı · olmaya · yaşadığı · konularda · yardımcı · olan hac · daha · iyi · bir · gelecek · oluşturuyoruz · inşa · etmekteyiz ·	TTE: 01'37"
QA	Human (3) Terminology and translation consistency: [Major] Style (readability, consistent style and tone): [Minor] Language quality (grammar, punctuation, spelling): [Minor]	

He also revised the sentences using the passive voice in the active voice if it fitted accordingly, as can be seen below:

Figure 13: Revising the Passive with an Active Voice

Source	We assess the importance and impact of the topics for the company, environment and society, and also for our stakeholders.	Words:	20
Suggestion			
Translation	Şirket, ortam ve toplum konularının önemli ve etkisini ve ayrıca paydaşlarımız için önemli ölçüde değerlendirilmektedir.	TTE:	19"
Revision	Şirket, ortam ve toplum konularının önemli ve etkisini ve ayrıca Konuların şirket, çevre, toplum ve paydaşlarımız için önemli ölçüde ve etkisini değerlendirilmektedir.	TTE:	02'52"
QA	Human (3)	Language quality (grammar, punctuation, spelling): [Minor] Style (readability, consistent style and tone): [Minor] Translation errors (mistranslation, additions or omissions): [Minor]	

All in all, we can argue that the fourth hypothesis, which stated *translators will not get primed by the output used in Task 1* because they would be more attuned to the specific types of errors primed in Task1, is confirmed.



CHAPTER 6 CONCLUSION

The present study offers a comprehensive survey of literature on MT and content production-related competences theorized in Translation Studies. While surveying the competence-centered studies in detail, it analyzes more recent work, including the four-dimensional multi-vortex competence model developed by Donald Kiraly, which describes each competence as a separate vortex, the study explains all the arguments made concerning the use of MT in Translation Studies and discussions thereof, i.e., MT-resistant skills, as well as the results of more concrete task-based studies on post-editing skills. Connected to this specific point of research, it peruses the data published from job market surveys by OECD and articulates on what future competences might need to be developed in the coming years.

Chapter 3 covers the vast array of research done on the subject of “post-editing”, elaborating on the dimensions of quality in MT output including the impact of expertise and experience on quality evaluations, domain, scope of evaluation, i.e., sentence or document-level, and the metrics used for automatic and manual evaluation. It also discusses the arguments made for human parity and delves into details of how human parity might be interpreted from a usability perspective involving divergent groups of users such as participants in crowdsourcing projects, seasoned experts and newcomers to the NLP and MT fields, from both translation-related and computer-science disciplines.

Chapter 4 articulates on the less-studied topic of the priming effect in the context of MT by first explaining the original study done by Kathrin Bock (1986). Therefore, combining the experimental setup of this subject matter with MT output evaluation, the present study conducted an experiment based on the Task-Based Learning (TBL) approach. The experimental setup, the participant profile and the results obtained from the study as well as the analysis based on these data are presented in Chapter 5. The results indicate that the first hypothesis, suggesting that *the post-editing by less experienced translators will be faster compared to professional translators because professionals would be attuned to errors and would need a longer period of post-editing time*, is not confirmed. As for the second hypothesis, suggesting *the number of total errors will be higher in the test set than the*

priming set, is confirmed. Connected to this hypothesis, for verifying the third hypothesis, stating *the number of errors annotated in the test set will be less in the issue types annotated in the priming set*, has proved to be erroneous and produced the opposite result, meaning *the number of errors annotated in the test set was greater in the issue types annotated in the priming set.*

As for the fourth hypothesis, which states *translators will spend more time on post-editing in Task 2*, was confirmed, as the data on time spent on PE clearly demonstrates. Furthermore, the second half of the hypothesis, suggesting *translators will not become primed by the output used in Task 1* because they would be more attuned to the specific types of errors primed in Task1, is also confirmed.

In terms of the error categories, Style and Translation Errors were the two error issue types under which the highest and most severe errors were annotated, while Language Quality and Style, specifically the consistency of Style, are the categories with the highest variance in both priming and test sets, indicating significant differences in terms of error annotations among participants in both sets. This specific result also demonstrates that an inverse priming effect (c.f. Ferreira and Bock 2006) where participants become more attentive to errors caused, they think, by the MT system was observed. This study reports on document-level evaluation scores, which allow comparing the qualities, PE effort and time spent on PE for the NMT output, i.e., Google NMT and Kantan AI MT, for two sustainability reports translated from English to Turkish. As for the limitations of the study, we can count three aspects, namely, the span of document-level evaluation, the number of participants and the research conducted for specific segments where the participants spend more or less time post-editing the errors and attempt to find a corresponding target language equivalent online. Though the experimental study presented here asked the participants to record their screen during both the PE tasks, not all were successful. As for the span specified for the document-level evaluation here, we limited it to 3 segments consisting of a minimum of 3 sentences up to a maximum of 20 without a break between them. Due to these technical hardships observed during the process, the number of participants eventually decreased to 7 in Task 2.

Further studies can explore the priming effect in a MT post-editing context with more participants. Moreover, why and how research behavior during the translation of terms and corporate styles changes depending on the experience of the translators might be investigated, as well as the specific length of the time period where participants continue to be more cautious about/attentive to repeating MT errors caused by the system, can be more precisely measured.



REFERENCES

- Akalın, Rahman, and Mehmet Gündoğdu. “Akademik Çeviri Eğitiminin Temel İlkeleri Üzerine Düşünceler: Uygulanan Ders İzlemleri Bağlamında Hedefler ve Beklentiler.” [Views on the basic principles of academic translation education: Goals and expectations from the point of applied curricula.] *Mersin University Journal of the Faculty of Education* 6, no.1 (2010): 79–93. <https://dergipark.org.tr/tr/download/article-file/160763>.
- Allen, Jeffrey. “Post-Editing.” In *Computers and Translation: A Translator’s Guide*, edited by Harold Somers. John Benjamins, 2003.
- Amidei, Jacopo, Paul Piwek and Alistair Willis. “Agreement is Overrated: A Plea for Correlation to Assess Human Evaluation Reliability.” In *Proceedings of the 12th International Conference on Natural Language Generation*, 2019. Doi: 10.18653/v1/W19-8642.
- Anderson, J. R. *The Architecture of Cognition*. Harvard Univ. Press, 1983.
- Balkul, Halil İbrahim. “Türkiye’de Akademik Çeviri Eğitiminde Çeviri Teknolojilerinin Yerinin Sorgulanması: Müfredat Analizi ve Öğretim Elemanlarının Konuya İlişkin Görüşleri Üzerinden Bir İnceleme.” [Questioning the place of translation technologies in Turkish academic translation education: An inquiry based on curriculum analysis and opinions of academics related to the topic.] PhD diss., Sakarya University, 2015.
- Banerjee, Satyanjeev and Alon Lavie. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.” In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, 2005.
- Bangalore, Srinivas, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. “Syntactic Variance and Priming Effects in Translation”. In *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, edited by Michael Carl, Srinivas Bangalore and Moritz Schaeffer. Springer, 2016.
- Barraut, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. “Findings of the 2020 Conference on Machine Translation

- (WMT20).” In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics, 2020.
- Barraut, Loïc Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. “Findings of the 2019 Conference on Machine Translation (WMT19).” In *Proceedings of the Fourth Conference on Machine Translation (WMT 19)*, 2019.
- Bayerl, S. and Paul, K.I. “What Determines Inter-Coder Agreement in Manual Annotations? A Meta-analytic Investigation.” *Computational Linguistics* 37, no.4(2011): 699–725. https://doi.org/10.1162/COLI_a_00074.
- Bengi-Öner, Işın, and Ülker İnce. 1995. “Çeviri Eğitiminde Kaynak Metni Yorumlamaya ve Çeviri Metni Oluşturmaya Yönelik Yöntem Arayışları: Geniş Bir Bakış Açısı.” [A search for methods to interpret source texts and produce target texts in translation training: A broader perspective.] *Çeviribilim: Çeviri Eğitiminde Yöntem Önerileri* no. 1, (1995): 125–142.
- Berman, A. *La traduzione e la lettera o L'albergo nella lontananza*. Quodlibet, 2003.
- Birkan Baydan, Esra. “Çeviri Eğitiminde Çeviri/Çevirmenlik Edinci: Problem Çözme ve Karar Verme Konusunda Bir Farkındalık Uygulaması.” [Translation/Translator Competence in Translator Training.] *Istanbul University Journal of Translation Studies* 4, no. 7 (2013): 103–125. <https://dergipark.org.tr/tr/download/article-file/157322>.
- . “Yapısalcılık-Sonrası Kuramsal Görüşler Doğrultusunda Çeviri Eğitimine Giriş: Toplumsal Oluşumcu Yaklaşım Işığında Uygulama Örnekleri.” [Introduction to translation through poststructuralist thinking in translator training: Social constructivist approach classroom activities.] PhD diss., Yıldız Technical University, 2011.
- Bobbitt, Zach. “How to Interpret the Intercept in a Regression Model (With Examples).” Last modified June 22, 2021. <https://www.statology.org/intercept-in-regression/>.
- . “One-Way ANOVA: Definition, Formula, and Example.” Last modified December 27, 2018. <https://www.statology.org/one-way-anova/>.
- Bock, J.Kathryn. “Syntactic Persistence in Language Production.” *Cognitive Psychology* 18, no. 3(1986): 355-387. [https://doi.org/10.1016/0010-0285\(86\)90004-6](https://doi.org/10.1016/0010-0285(86)90004-6).
- Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. “Findings

- of the 2013 Workshop on Statistical Machine Translation.” In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, (2013): 1–44. Association for Computational Linguistics. <http://www.statmt.org/wmt13/pdf/WMT01.pdf>.
- Branigan, Holly P., Martin J. Pickering, Jamie Pearson, and Janet F. McLean. “Linguistic Alignment between People and Computers.” *Journal of Pragmatics*, no. 42 (2010): 2355-2368. <https://doi.org/10.1016/j.pragma.2009.12.012>.
- Brown, Keith and Jim Miller. *The Cambridge Dictionary of Linguistics*. Cambridge University Press, 2013. <https://doi.org/10.1017/CBO9781139049412>.
- Burchardt, A., V. Macketanz, J. Dehdari, G. Heigold, J.-T. Peter, and P. Williams. “A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines.” *The Prague Bulletin of Mathematical Linguistics* 108 (2017): 159–170. doi:10.1515/pralin-2017-0017.
- Carl, Michael and Moritz Schaeffer. “Measuring Translation Literality.” In *Translation in Transition: Between Cognition, Computing and Technology*, edited by Arnt Lykke Jakobsen and Bartolomé Mesa-Lao. John Benjamins, 2017.
- Carl M., S. Gutermuth, S. Hansen-Schirra. “Post-Editing Machine Translation: A Usability Test For Professional Translation Settings.” In *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*, edited by Ferreira A, Schwieter JW. John Benjamins, 2015.
- Calvo, Elisa. “Translation and/or Translator Skills as Organising Principles for Curriculum Development Practice.” *The Journal of Specialized Translation* 16, no. 16 (July 2011): 5-25, <https://doi.org/10.26034/cm.jostrans.2011.484>.
- Castilho, Sheila. “Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation.” In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 34–45, Online. Association for Computational Linguistics, 2021.
- . “Measuring Acceptability of Machine Translated Enterprise Content.” PhD diss., Dublin City University, 2016.
- Castilho, S., O’Brien S, Alves F. and O’Brien M. “Does Post-Editing Increase Usability? A Study with Brazilian Portuguese as Target Language.” In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation*, 2014.

- Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. “Approaches to Human and Machine Translation Quality Assessment.” In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Springer, 2018.
- Castilho, Sheila and Rebecca Knowles. “A Survey of Context in Neural Machine Translation and Its Evaluation.” *Natural Language Processing*. (2024):1-31. doi:10.1017/nlp.2024.7.
- Castilho, Sheila, J. Moorkens, F. Gaspari, R. Sennrich, V. Sosoni, P. Georgakopoulou, P. Lohar, A. Way, A. Valerio Miceli Barone, and M. Gialama. “A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators.” In *Proceedings of MT Summit 2017* (2017a): 116-131. Asia-Pacific Association for Machine Translation, <https://aclanthology.org/2017.mtsummit-papers.10/>.
- Castilho, S., J. Moorkens, F. Gaspari, I. Calixto, J. Tinsley, and A. Way. “Is Neural Machine Translation the New State-of-The-Art?” *The Prague Bulletin of Mathematical Linguistics* 108 (June 2017b): 109–120. doi:10.1515/pralin-2017-0013.
- Castilho, Sheila, Maja Popović, and Andy Way. “On Context Span Needed for Machine Translation Evaluation.” In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC’20)*, 2020. <http://www.lrec-conf.org/proceedings/lrec2020/index.html>.
- Chaouch Orozco, Adel, Gonzáles Alonso, Jorge; Duñabeitia, Jon Andoni; Rothman, Jason. “Are Translation Equivalents Really Equivalent? Evidence from Concreteness Effects in Translation Priming.” *International Journal of Bilingualism* 28, no. 2 (2023). <https://doi.org/10.1177/13670069221146641>.
- Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches.” In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, 2014.
- Comparin, Lucia and Sara Mendes. “Using Error Annotation to Evaluate Machine Translation and Human Post-Editing in a Business Environment.” In *Proceedings of EAMT 2017*. Accessed May 1, 2025. https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017_paper_76.pdf.
- Costa-jussà, Marta R. and Mireia Farrús. “Towards Human Linguistic Machine Translation Evaluation.” *Digital Scholarship in the Humanities* 30, no. 2 (2015):157-166. doi:10.1093/llc/fqt065.

- Cowan, Benjamin R., Holly P. Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. "Voice Anthropomorphism, Interlocutor Modelling and Alignment Effects on Syntactic Choices in Human Computer Dialogue." *International Journal of Human-Computer Studies* 83, (2015): 27–42. <https://doi.org/10.1016/j.ijhcs.2015.05.008>.
- Cumbreño, Cristina and Nora Aranberri. "What Do You Say? Comparison of Metrics for Post-editing Effort." In *Explorations in Empirical Translation Process Research*, edited by Michael Car. Springer, 2021.
- Çetiner, Caner. "Redefining the Competences of Translators in Light of the Increasing Use of Post-Editing." In *Translation and Interdisciplinarity Diverse Contacts, Relationships and Effects*, edited by Faruk Yücel, Mehmet Tahir Öncü. Logos Verlag, 2022.
- . "Sustainability of Translation as a Profession: Changing Roles of Translators in light of the Developments in Machine Translation Systems." *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi* Ö9 (2021): 575-586. <https://doi.org/10.29000/rumelide.985014>.
- . "Undergraduate Level Translation Students' Attitudes towards Machine Translation Post-Editing Training." *International Journal of Languages' Education and Teaching* 7, no.1 (2019): 110-120. Doi: 10.18298/ijlet.3242.
- Daems, Joke, Michael Carl, Sonia Vandepitte, Robert Hartsuiker and Lieve Macken. "The Effectiveness of Consulting External Resources During Translation and Post-editing of General Text Types." In *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, edited by Michael Carl, Srinivas Bangalore and Moritz Schaeffer. Springer, 2016.
- Daems, J., and Macken, L. "Post-Editing Human Translations and Revising Machine Translations: Impact on Efficiency and Quality." In *Translation Revision and Post-Editing: Industry Practices and Cognitive Processes*, edited by Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera. Routledge, 2021.
- . "Interactive Adaptive SMT versus Interactive Adaptive NMT: A User Experience Evaluation." *Machine Translation* 33, no. 1-2 (2019). <https://doi.org/10.1007/s10590-019-09230-z>.
- Daems J., S. Vandepitte, R. Hartsuiker and L. Macken. "The Impact of Machine Translation Error Types on Post-Editing Effort Indicators." In *Proceedings of the 4th Workshop on Post-Editing Technology and Practice*, 2015.
- Davis, Chris, Rosa Sánchez-Casas, José E. García-Albea, Marc Guasch, Margarita Molero and Pilar Ferré. "Masked Translation Priming: Varying Language Experience and Word Type with Spanish-English Bilinguals." *Bilingualism: Language and Cognition* 13, no. 2 (2010): 137–155. doi:10.1017/S1366728909990393.

- de Almeida, Giselle. “*Translating the Post-Editor: An Investigation of Post-Editing Changes and Correlations with Professional Experience Across Two Romance Languages.*” PhD diss., Dublin City University, 2013.
https://doras.dcu.ie/view/people/de_Almeida=3AGiselle=3A=3A.html.
- de Almeida, G., and S. O’Brien. “Analysing Post-Editing Performance: Correlations with Years of Translation, Experience.” In *Proceedings of 14th Annual Conference of the European Association for Machine Translation*, 2010. Available from: <http://mt-archive.info/EAMT-2010-Almeida.pdf>.
- De Beaugrande R., and Dressler W. *Introduction to Text Linguistics*. Longman, 1981.
- Dell, Gary S. and Cassandra L. Jacobs. “Successful Speaking: Cognitive Mechanisms of Adaptation in Language Production.” In *Neurobiology of Language*, edited by Gregory Hickok, Steven L. Small. Academic Press, 2016.
- Densmer, L. “Light and Full MT Post-Editing Explained.” (2014).
<http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained>.
- DePalma D. and Kelly N. “The Business Case for Machine Translation.” In *Proceedings of Machine Translation Summit XII: Tutorials*, 2009. <https://aclanthology.org/www.mt-archive.info/MTS-2009-DePalma-ppt.pdf>.
- Depraetere, Ilse. “What Counts As Useful Advice in a University Post-Editing Training Context? Report On a Case.” In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 2010.
- Depraetere, Ilse, Nathalie de Sutter, Arda Tezcan. “Post-edited Quality, Post-editing Behaviour and Human Evaluation: A Case Study.” In *Post-editing of Machine Translation: Processes and Applications*, edited by Sharon O’Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia. Cambridge Scholars Publishing, 2014.
- de Almeida, Giselle. “Translating the Post-Editor: An Investigation of Post-Editing Changes and Correlations with Professional Experience across Two Romance Languages.” PhD diss., Dublin City University, 2013.
- de Almeida, Giselle and Sharon O’Brien. “Analysing Post-Editing Performance: Correlations with Years of Translation Experience.” In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, edited by François Yvon and Viggo Hansen. European Association for Machine Translation, 2010.
<https://aclanthology.org/2010.eamt-1.19/>.

- do Carmo, Félix, Dimitar Shterionov, Joss Moorkens, Joachim Wagner, Murhaf Hossari, Eric Paquin, Dag Schmidtke, Declan Groves and Andy Way. “A Review of the State-of-the-Art in Automatic Post-Editing.” *Machine Translation* 35 (2021): 101–143, <https://doi.org/10.1007/s10590-020-09252-y>.
- Doherty, S. “Issues in Human and Automatic Translation Quality Assessment.” In *Human Issues in Translation Technology*, edited by Kenny, D. Routledge, 2017.
- . “The Impact of Translation Technologies on the Process and Product of Translation.” *International Journal of Communication* 10 (2016): 947–969.
- Doherty, S. and Sharon O’Brien. “Assessing the Usability of Raw Machine Translated Output: A User-Centred Study Using Eye Tracking.” *International Journal of Human-Computer Interaction* 30, no.1 (2014):40–51. <https://doi.org/10.1080/10447318.2013.802199>.
- Drugan, J. *Quality in Professional Translation: Assessment and Improvement*. Bloomsbury, 2013.
- Dybkjær L., Bernsen N., Minker W. “Evaluation and Usability of Multimodal Spoken Language Dialogue Systems.” *Speech Communication* 43, no.1-2 (2004): 33–54. <https://doi.org/10.1016/j.specom.2004.02.001>.
- European Commission. *EMT Competence Framework 2022*. Accessed 25 April 2025. Updated version. https://commission.europa.eu/system/files/2022-11/emt_competence_fw_2022_en.pdf.
- ELIA, EMT, EUATC, FIT Europe, GALA, LIND, and Women in Localization. *European Language Industry Survey 2022. Trends, Expectations and Concerns of the European Language Industry*. Accessed May 1, 2025. https://fit-europe-rc.org/wp-content/uploads/2022/03/ELIS-2022_survey_results_final_report.pdf.
- EMT Network. *European Master’s in Translation Competence Framework 2017*. Accessed 25 April 2025. https://commission.europa.eu/system/files/2018-02/emt_competence_fw_2017_en_web.pdf.
- Ersoy, Hüseyin, and Halil İbrahim Balkul. “Teknolojik Gelişmelerin Çevirmen ve Çeviri Mesleği Açısından Olumlu ve Olumsuz Etkileri: Çeviri Alanında Yeni Yaklaşımlar.” [The positive and negative effects of technological developments from the perspective of translator and translation profession: New approaches in translation field.] *Journal of Academic Inquiries* 7, no. 2 (2012): 295–307. <https://dergipark.org.tr/tr/download/article-file/17758>.
- Esen-Eruz, Sâkine. *Akademik Çeviri Eğitimi: Çeviri Amaçlı Metin Çözümlemesi* [Academic translator training: Translation-oriented text analysis]. Multilingual, 2008.

- Esperança-Rodier, E., C. Rossi, A. Bérard, and L. Besacier. "Evaluation of NMT and SMT Systems: A Study on Uses and Perceptions." In *Proceedings of the 39th conference Translating and the Computer*. AsLing, 2017. <https://shs.hal.science/halshs-01970463/document>.
- Faigley, L. and S. Witte. "Analyzing Revision." *College Composition and Communication* 32, no. 4 (1981): 400-414. <https://doi.org/10.2307/356602>.
- Federico, M., A. Cattelan and M. Trombetti. "Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation." In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012)*, 2012.
- Ferreira, V. S. "Sentence Production." In *Encyclopedia of Neuroscience*. Academic Press, 2009. <https://doi.org/10.1016/B978-008045046-9.01885-4>.
- Ferreira, Victor S. and Kathryn Bock. "The Functions of Structural Priming." *Language and Cognitive Processes* 21, no. 7-8 (2006): 1011-1029. <https://doi.org/10.1080/01690960600824609>.
- Fiederer, R. and S. O'Brien. "Quality and Machine Translation: A Realistic Objective?" *Journal of Specialised Translation* 11 (2009): 52-72.
- Fields P, D. Hague, G.S. Koby, A. Lommel, A. Melby. "What is Quality? A Management Discipline and the Translation Industry Get Acquainted." *Revista Tradumàtica* 12(2014):404-412. Accessed May 1, 2025. <https://revistes.uab.cat/tradumatica/article/view/n12-fields-hague-koby-et-al>.
- Flanagan, Marian and Tina Paulsen Christensen. "Testing Post-Editing Guidelines: How Translation Trainees Interpret Them and How to Tailor Them for Translator Training Purposes." *The Interpreter and Translator Trainer* 8, no. 2 (2014): 257-275. <https://doi.org/10.1080/1750399X.2014.936111>.
- Forcada, Mikel L. "Making Sense of Neural Machine Translation." *Translation Spaces* 6, no. 2 (2017): 291-309. <https://doi.org/10.1075/ts.6.2.06for>.
- . "On the Annotation of TMX Translation Memories for Advanced Leveraging in Computer-Aided Translation." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*: 4374-4378. European Language Resources Association (ELRA), 2014.
- Gao M, W. Xu, C. Callison-Burch. "Cost Optimization in Crowdsourcing Translation: Low Cost Translations Made even Cheaper." In *Proceedings of the 2015 Conference of the*

- North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2015.
- Garcia, Ignacio. “Translating by Post-Editing: Is It a Way Forward?” *Machine Translation* 25, no.3 (2011): 217–237. <https://doi.org/10.1007/s10590-011-9115-8>.
- . “The Proper Place of Professionals (and Non-Professionals and Machines) in Web Translation.” *Tradumàtica* 8 (2010): 1-7. DOI: 10.5565/rev/tradumatica.98.
- Garvin, David A. “What does ‘Product Quality’ Really Mean?” *Sloan Management Review* (Fall 1984): 25-43. http://www.oqrm.org/English/What_does_product_quality_really_means.pdf.
- Gaspari, F. “Perception vs Reality: Measuring Machine Translation Post-Editing Productivity.” In *Proceedings of the Third Workshop on Post-Editing Technology and Practice at the 11th Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, 2014.
- Gile, D. *La traduction. La comprendre, l'apprendre*. Presses universitaires de France, 2005.
- Giménez J, L. Márquez, E. Comelles, I. Catellón, V. Arranz. “Document-Level Automatic MT Evaluation Based on Discourse Representations.” In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 2010.
- Global Reporting Initiative. *GRI Standards*. 2023. <https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/>.
- Goto, S., D. Lin, and T. Ishida. “Crowdsourcing for Evaluating Machine Translation Quality.” *LREC* (2014): 3456-3463. <https://aclanthology.org/www.mt-archive.info/10/LREC-2014-Goto.pdf>.
- Gouadec D. *Translation as a Profession*. John Benjamins, 2007.
- Göpferich, Susanne. “Translation Competence: Explaining Development and Stagnation from a Dynamic Systems Perspective.” *Target* 25, no. 1 (2013): 61–76. <https://doi.org/10.1075/target.25.1.06goe>.
- . “Towards a Model of Translation Competence and its Acquisition: The Longitudinal Study TransComp.” In *Behind the Mind: Methods, Models and Results in Translation Process Research*, edited by Susanne Göpferich, Arnt Lykke Jakobsen and Inger M. Mees. Samfundslitteratur, 2009.

- Guerberof Arenas, Ana. "The Role of Professional Experience in Post-Editing from a Quality and Productivity Perspective." In *Post-editing of Machine Translation: Processes and Applications*, edited by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia. Cambridge Scholars Publishing, 2014.
- . "Productivity and Quality in the Post-Editing of Outputs from Translation Memories and Machine Translation." PhD diss., Universitat Rovira i Virgili, 2012.
- . "Productivity and Quality in the Postediting of Outputs from Translation Memories and Machine Translation." *Localis Focus* 7, no.1 (2009a):11–21.
- . "Productivity and Quality in MT Post-Editing." In *Proceedings of MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. AMTA, 2009b.
- . "Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation." *The International Journal of Localisation*, no. 7 (2008): 11–21.
- Gutt, E.-A. "Applications of Relevance Theory to Translation: A Concise Overview." (2004). Accessed May 1, 2025. DOI:[10.13140/RG.2.2.33067.80162](https://doi.org/10.13140/RG.2.2.33067.80162).
- Guzmán F, S. Joty, L. Márquez, P. Nakov P. "Using Discourse Structure Improves Machine Translation Evaluation." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.
- Hansen, Gyde. "A Classification of Errors in Translation and Revision." In *CIUTI Forum 2008. Enhancing Translation Quality: Ways, Means, Methods*, edited by M. Forstner, H. Lee- Jahnke, and P. Schmit, 313–326. Peter Lang, 2009.
- Haro-Soler, Maria del Mar, and Don Kiraly. "Exploring Self-Efficacy Beliefs in Symbiotic Collaboration with Students: An Action Research Project." *The Interpreter and Translator Trainer* 13, no. 3 (2019): 255–270. <https://doi.org/10.1080/1750399X.2019.1656405>.
- He, Yifan and Way A. "Improving the Objective Function in Minimum Error Rate Training." In *Proceedings of Machine Translation Summit XII: Posters*, 2009a.
- . "Metric and Reference Factors in Minimum Error Rate Training." *Machine Translation* 24, no.1 (2009b): 27–38.
- Hedges, Chris. *Empire of Illusion: The End of Literacy and the Triumph of Spectacle*. Nation Books, 2009.

- Heyselaar, Evelien, Peter Hagoort, and Katrien Segaert. “How Social Opinion Influences Syntactic Processing—An Investigation Using Virtual Reality.” *PLOS ONE* 12, no. 4(2017a): e0174405. <https://doi.org/10.1371/journal.pone.0174405>.
- . “In Dialogue with an Avatar, Language Behavior is Identical to Dialogue with a Human Partner.” *Behavior Research Methods*, no. 49 (2017b): 46–60. <https://doi.org/10.3758/s13428-015-0688-7>.
- Horguelin, P. A., and L. Brunette. *Pratique de la révision*. Linguattech, 1998.
- Horváth, P. I. *A Szakfordítások Lektorálása. Elmélet És Gyakorlat* [Revising Specialised Translations. Theory and Practice]. Tinta Könyvkiadó, 2011.
- House, J. *Translation Quality Assessment: A Model Revisited*. Gunter Narr, 1997.
- Hovy E, King M, Popescu-Belis A. “Principles of Context-Based Machine Translation Evaluation.” *Machine Translation* 17, no.1(2002):43–75.
- Hu, Ke and Cadwell, Patrick. “A Comparative Study of Postediting Guidelines.” *Baltic Journal of Modern Computing* 4, no.2 (2016): 346-353.
- Hurtado Albir, A. “Competence-Based Curriculum Design for Training Translators.” *The Interpreter and Translator Trainer* 1, no. 2 (2007): 63–195. doi:10.1080/1750399X.2007.10798757.
- İnce, Ülker, and Işın Bengi-Öner. *Kızılcık Karpuz Olur mu Hiç? İlahi Çevirmen!* [How come the cornelian cherry turns into a watermelon? Ah translator!]. Diye, 2009.
- Inoue, Izumi and Christopher N. Candlin. “Applying Task-Based Learning to Translator Education: Assisting the Development of Novice Translators’ problem-solving expertise.” *Translation and Interpreting Studies* 10, no.1 (2015), 58–86. <https://doi.org/10.1075/tis.10.1.04ino>.
- ISO/TS 11669: Translation Projects — General Guidance. 2012.
- ISO/TR 16982: Ergonomics of Human-System Interaction — Usability Methods Supporting Human-Centred Design. 2002.
- ISO 17100: Translation Services — Requirements for Translation Services. 2015.
- ISO 18587: Translation services — Post-Editing of Machine Translation Output — Requirements. 2017.

ISO 30042: Management of Terminology Resources — TermBase eXchange (TBX). 2008.

Jakobsen, Arnt Lykke. “Moving Translation, Revision, and Post-Editing Boundaries.” In *Moving Boundaries in Translation Studies*, edited by Helle V. Dam, Matilde Nisbeth Brøgger and Karen Korning Zethsen. Routledge, 2019.

Jääskeläinen R. “Quality and Translation Process Research.” In *Reembedding Translation Process Research*, edited by Muñoz Martín R. John Benjamins, 2016.

Jiménez-Crespo, Miguel Angel. “Crowdsourcing and Translation Quality: Novel Approaches in the Language Industry and Translation Studies”. In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty. Springer, 2018.

———. “The Evaluation of Pragmatic and Functionalist Aspects in Localization: Towards a Holistic Approach to Quality Assurance”. *The Journal of Internationalization and Localization* 1 (2009): 60-93.

Kelly, Dorothy. “Mobility Programmes as a Learning Experience for Translation Students: Development and Assessment of Specific Translations and Transferable Generic Competences in Study Abroad Contexts.” In *Translator and Interpreter Training: Issues, Methods and Debates*, edited by John Kearns. Continuum, 2008.

———. *A Handbook for Translator Trainers: A Guide to Reflective Practice*. St Jerome, 2005.

Kenny, D. “Sustaining Disruption? The Transition From Statistical To Neural Machine Translation.” *Revista Tradumàtica. Tecnologies de la Traducció* 16 (2018): 59–70. <https://doi.org/10.5565/rev/tradumatica.221>.

Kenny, D. and Doherty, S. Statistical Machine Translation in the Translation Curriculum: Overcoming Obstacles and Empowering Translators. *The Interpreter and Translator Trainer* 8, no. 2 (2014): 276-294. <https://doi.org/10.1080/1750399X.2014.936112>.

Kiraly, D. “Occasioning Translator Competence: Moving beyond Social Constructivism Towards a Postmodern Alternative to Instructionism.” *Translation and Interpreting Studies* 10, no. 1 (2015): 8–32. <https://doi.org/10.1075/tis.10.1.02kir>.

———. “Towards A View of Translator Competence as an Emergent Phenomenon: Thinking Outside the Box(es) in Translator Education.” In *New Prospects and Perspectives for Educating Language Mediators*, edited by Kiraly, Don/Silvia Hansen-Schirra/Karin Maksymski. Narr Verlag, 2013.

- . “Beyond Social Constructivism: Complexity Theory and Translator Education.” *Translation and Interpreting Studies* 1, no. 1 (2006): 68–86. doi:10.1075/tis.1.1.05kir.
- . “Preparing Students Today to Meet Market Demands Tomorrow.” In *Internationales CIUTI-Forum: Marktorientierte Translationsausbildung* [International CIUTI-forum: Market-oriented translation training], edited by Martin Forstner and Hannelore Lee-Jahnke, 101–118. Peter Lang, 2004.
- . *A Social Constructivist Approach to Translator Education; Empowerment from Theory to Practice*. St. Jerome Publishing, 2000.
- . *Pathways to Translation: Pedagogy and Process*. The Kent State University Press, 1995.
- Koby G.S., Fields P, Hague D, Lommel A, Melby A. “Defining Translation Quality.” *Revista Tradumàtica* 12 (2014): 413–420. Accessed May 1, 2025. https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p413.pdf.
- Koehn, P. “Enabling Monolingual Translators: Post-Editing Vs. Options.” In *Proceedings Of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2010.
- Koehn, P., and Knowles, R. “Six Challenges for Neural Machine Translation.” In *Proceedings of the First Workshop on Neural Machine Translation*, edited by T. Luong, A. Birch, G. Neubig, and A. Finch. Association for Computational Linguistics, 2017. <https://doi.org/10.48550/arXiv.1706.03872>.
- . “Europarl: A Parallel Corpus for Statistical Machine Translation.” In *Proceedings of MT Summit X*, 2005.
- Konttinen, Kalle, Leena Salmi and Maarit Koponen. “Revision and Post-Editing Competences in Translator Education.” In *Translation Revision and Post-Editing: Industry Practices and Cognitive Processes*, edited by Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera. Routledge, 2021.
- Koponen, M. “How to Teach Machine Translation Post-Editing? Experiences from a Post-Editing Course.” In *Proceedings of the 4th Workshop on Post-editing Technology and Practice*, 2015.
- . “Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations.” In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 2012.

- Koponen, M. and Salmi, L. “On the Correctness of Machine Translation: A Machine Translation Post-editing Task.” *The Journal of Specialized Translation* 23 (2015). https://jostrans.soap2.ch/issue23/art_koponen.php.
- Kornfilt, Jaklin. *Turkish*. Routledge, 1997.
- Kosmaczewska, Kasia and Matt Train. “Application of Post-Edited Machine Translation in Fashion eCommerce.” In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. European Association for Machine Translation, 2019.
- Krings, Hans P. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press, 2001.
- Kurultay, Turgay. “Neler Olmazsa Çeviri Eğitimi Olmaz? Çeviri Eğitiminde Temel İlkeler Belirlemeye Yönelik Bir Deneme.” [What are the prerequisites of translator training? An attempt to determine the basic principles of translator training.] In *Forum: Türkiye’de Çeviri Eğitimi; Nereden Nereye?* [Forum: Translator training in Turkey; From where to where?], edited by Turgay Kurultay and İlknur Birkandan. Sel, 1997.
- Künzli, Alexander. “Teaching and Learning Translation Revision: Some Suggestions based on Evidence from a Think-Aloud Protocol Study.” In *Current Trends in Translation Teaching and Learning*, edited by Mike Garant. Helsinki University Press, 2006.
- Labaka, G., España-Bonet C, Marquez L, Sarasola K. “A Hybrid Machine Translation Architecture Guided by Syntax.” *Machine Translation* 28, (2014): 91–125. <https://doi.org/10.1007/s10590-014-9153-0>.
- Lacruz, I. and Shreve G M. “Pauses and Cognitive Effort in Post-editing.” In *Post-editing of Machine Translation: Processes and Applications*, edited by S O’Brien S, LW Balling, M Carl, M Simard, L Specia. Cambridge Scholars Publishing, 2014.
- Landis, J.R. and Koch G G. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1977): 157–74.
- Lavie, Alon, Alok Parlikar and Vamshi Ambati. 2008. “Syntax-driven Learning of Sub-sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora.” In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, 2008.
- Läubli, S., Amrhein, C., Düggelein, P., Gonzalez, B., Zwahlen, A. and Volk, M. “Post- Editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain.” In *Proceedings of MT Summit XVII*.

Volume 1: Research Track. European Association for Machine Translation (EAMT), 2019.

- Läubli, S., Sennrich, R., and Volk, M. “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)*. Association for Computational Linguistics, 2018. Doi: 10.18653/v1/D18-1512.
- Lehmann, Sabine, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnon, Judith Baur, Lorna Balkan, and Doug Arnold. “TSNLP - Test Suites for Natural Language Processing.” In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.
- Lemhöfer, K. and Broersma M. “Introducing LexTALE: A Quick and Valid Lexical Test for Advanced Learners of English.” *Behavior Research Methods* 44, (2012): 325-43. doi: 10.3758/s13428-011-0146-0.
- Lesznyák, M. “Conceptualizing Translation Competence.” *Across Languages and Cultures* 8, no. 2 (2007): 167–194. doi:10.1556/Acr.8.2007.2.2.
- Lommel, Arle. “Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies.” In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Springer, 2018.
- Lommel, A., Popović, M. and Burchardt, A. “Assessing Inter-Annotator Agreement for Translation Error Annotation.” In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. International Conference on Language Resources and Evaluation (LREC-14)*, 2014. https://mte2014.github.io/Talks/Session1/08_Maja_slides-iaa.pdf.
- Lommel, A., Uszkoreit, H., Burchardt A. “Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics.” *Revista Tradumàtica* 12 (2014): 455–463. https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p455.pdf. Accessed May 1, 2025.
- Macken, Lieve, Daniel Prou and Arda Tezcan. “Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process.” *INFORMATICS-BASEL* 7, no.2 (2020). doi:10.3390/informatics7020012.
- Marconi, G., L. Vergolini and F. Borgonovi. “The Demand for Language Skills in the European Labour Market: Evidence from Online Job Vacancies.” In *OECD Social*,

- Employment and Migration Working Papers no.294*. OECD Publishing, 2023.
<https://doi.org/10.1787/e1a5abe0-en>.
- Marg, L. “The Trials and Tribulations of Predicting Post-Editing Productivity.” In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. European Language Resources Association (ELRA), 2016.
- Martínez Mateo, R. “A Deeper Look into Metrics for Translation Quality Assessment (TQA): A Case Study.” *Miscelánea: A Journal of English and American Studies* 49 (2014), 73-93. https://doi.org/10.26754/ojs_misc/mj.20148792.
- Mauriello, G. “Teacher’s Tools in a Translation Class.” In *Teaching Translation and Interpreting: Training Talent and Experience. Papers from the First Language International Conference, Elsinore, Denmark, 1991*, edited by C. Dollerup and A. Loddegaard. John Benjamins, 1992.
- Mellinger, Christopher D. and Gregory M. Shreve. “Match Evaluation and Over-Editing in a Translation Memory Environment.” In *Reembedding Translation Process Research*, edited by Munoz Martín. John Benjamins 2016.
- Mesa-Lao, B. “Introduction to Post-editing – The CasMaCat GUI.” (2013.) http://bridge.cbs.dk/projects/seecat/material/hand-out_post-editing_bmesa-lao.pdf
- Metametrics. “Lexile.” About Lexile® Measures for Reading. Accessed May 1, 2025. <https://metametricsinc.com/parents-and-students/lexile-for-parents-and-students/lexile-for-reading/>.
- Mitchell, Linda. “The Potential and Limits of Lay Post-Editing in an Online Community.” In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 2015.
- Moniz, Helena, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, Margot Fonteyne, eds. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 2022.
- Moorkens, Joss. “What to Expect from Neural Machine Translation: A Practical In-Class Translation Evaluation Exercise.” *The Interpreter and Translator Trainer* 12, no.4 (2018): 375–387. doi:10.1080/1750399X.2018.1501639.

- . “Under Pressure: Translation in times of Austerity.” *Perspectives* 25, no. 3 (2017): 464–477. <https://doi.org/10.1080/0907676X.2017.1285331>.
- Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty. “Introduction.” In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Springer, 2018.
- Moorkens, Joss, Andy Way, and Séamus Lankford. *Automating Translation*. Routledge, 2025.
- Moorkens, Joss and Sharon O’Brien. “Assessing User Interface Needs of Post-Editors of Machine Translation.” In *Human Issues in Translation Technology* (The IATIS Yearbook), edited by Dorothy Kenny. Routledge, 2017.
- Moorkens, Joss, O’Brien, S., da Silva I. A. L., de Lima Fonseca N. B., and Alves F. “Correlations of Perceived Post-Editing Effort with Measurements of Actual Effort.” *Machine Translation* 29, no. 3 (2015): 267–284.
- Mossop, Brian. *Revising and Editing for Translators*. 4th ed. Routledge, 2020/2014/2007/2001.
- MQM. “MQM (Multidimensional Quality Metrics).”. Accessed May 1, 2025. <https://themqm.org>.
- National Accreditation Authority for Translators and Interpreters. “NAATI Certification System.” Accessed May 1, 2025. <https://www.naati.com.au/certification/>.
- Neubert, Albrecht. “Competence in Language, in Languages, and in Translation.” In *Developing Translation Competence*, edited by Christina Schäffner and Beverly Adab. John Benjamins, 2000.
- Neubig, Graham, Makoto Morishita, and Satoshi Nakamura. “Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015.” In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*. Workshop on Asian Translation, 2015.
- Nida, Eugene. *Towards a Science of Translation*. Brill, 1964.
- Nida, E. and Taber C. R. *The Theory and Practice of Translation*. Brill, 1969.
- Nielsen, J. *Usability Engineering*. Morgan Kaufmann, 1993.

- Nitzke, Jean and Anne-Kathrin Gros. 2021. "Referential Changes in Revision and Post-Editing." In *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*, edited by Maarit Koponen, Brian Mossop, Isabelle S. Robert, Giovanna Scocchera. Routledge, 2021.
- Nitzke, Jean and Silvia Hansen-Schirra. *A Short Guide to Post-Editing*. Language Science Press, 2021.
- Nord, Christiane. 1997. *Translation as a Purposeful Activity*. St. Jerome, 1997.
- Norton, Scott. *Developmental Editing: A Handbook for Freelancers, Authors, and Publishers*. University of Chicago Press, 2009.
- O'Brien, Sharon. "Towards a Dynamic Quality Evaluation Model for Translation." *The Journal of Specialised Translation* 17(2012): 55–77. https://jostrans.soap2.ch/issue17/art_obrien.pdf.
- . 2011. "Towards Predicting Post-Editing Productivity." *Machine Translation* (25): 197-215. <https://doi.org/10.1007/s10590-011-9096-7>.
- . "Eye-Tracking and Translation Memory Matches." *Perspectives* 14, no.3 (2007):185–204. <https://doi.org/10.1080/09076760708669037>.
- . "Teaching Post-Editing: A Proposal For Course Content." In *Proceedings of the 6th EAMT Workshop on "Teaching Machine Translation"*. European Association for Machine Translation, 2002.
- O'Brien S., Balling, L.W., Carl M., Simard M., Specia L., eds. *Post-Editing of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing, 2014.
- O'Brien, Sharon, Johann Roturier, and Giselle de Almeida. "Postediting Machine Translation Output Guidelines." In *Proceedings of Machine Translation Summit XII: Tutorials*, 2009.
- O'Brien S., Simard, M., Specia L., eds. *Workshop on Post-Editing Technology and Practice (WPTP 2013)*. Machine Translation Summit XIV, 2013.
- O'Brien S., Simard, M., Specia L., eds. *Workshop on Post-Editing Technology and Practice (WPTP 2012)*. Association for Machine Translation in the Americas, 2012.
- Onishi, Nanami and Yamada Masaru. "Why Translator Competence in Information Searching Matters: An Empirical Investigation into Differences in Searching Behavior Between Professionals and Novice Translators." *Invitation to Interpreting and*

- Translation Studies* 22, (2020):1-23. Japan Association for Interpreting and Translation Studies.
- Öner, Işın and Senem Öner Bulut. “Post-Editing Oriented Human Quality Evaluation of Neural Machine Translation in Translator Training: A Study on Perceived Difficulties and Benefits.” *transLogos* 4, no.1 (2021): 100-124. <https://dx.doi.org/10.29228/transLogos.33>.
- Öner Bulut, Senem. “Integrating Machine Translation into Translator Training: Towards ‘Human Translator Competence’?”. *TransLogos Translation Studies Journal* 2, no. 2 (2019): 1-26. <https://doi.org/10.29228/transLogos.11>.
- Öner, Senem. “Translator as Researcher: Perspectives on Training and Life-Long Professional Improvement of Translators.” *Procedia* 106 (2013): 3159–3166. <https://doi.org/10.1016/j.sbspro.2013.12.365>.
- Öner Bulut, Senem and Nilüfer Alimen. 2023. "Translator Education as a Collaborative Quest for Insights into the Re-positioning of the Human Translator (Educator) in the Age of Machine Translation: the Results of a Learning Experiment." *Interpreter and Translator Trainer* 17, no.3 (2023): 375-392. <https://doi.org/10.1080/1750399x.2023.2237837>.
- Öztürk Baydere, Hilal. “Makine Çevirisi ve Dijitalleşme Bağlamında Çevirmenlik Mesleğinin Dönüşümü [Transformation of the Translation Profession within the Context of Machine Translation and Digitalization].” PhD diss., İstanbul Üniversitesi, 2022.
- PACTE. “First Results of PACTE Group’s Experimental Research on Translation Competence Acquisition: The Acquisition of Declarative Knowledge of Translation.” In *Minding translation/Con la traducción en mente*, edited by R. M. Martín. Universitat d’Alacant, 2014.
- . “Results of the Validation of the PACTE Translation Competence Model: Translation Problems and Translation Competence.” In *Methods and Strategies of Process Research: Integrative Approaches in Translation Studies*, edited by C. Alvstad, A. Hild, and E. Tiselius, 317–341. Benjamins, 2011.
- . “First Results of a Translation Competence Experiment: ‘Knowledge of Translation’ and ‘Efficacy of the Translation Process’.” In *Translator and Interpreter Training: Issues, Methods and Debates*, edited by J. Kearns. Continuum, 2008.
- . “Investigating Translation Competence: Conceptual and Methodological Issues.” *Meta* no. 50 (2005): 609-619. <https://doi.org/10.7202/011004ar>.

- . “Building a Translation Competence Model.” In *Triangulating Translation: Perspectives in Process Oriented Research*, edited by F. Alves. Benjamins, 2003.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (July 2002): 311-318. <http://www.aclweb.org/anthology/P02-1040.pdf>. Accessed May 1, 2025.
- Parra Escartín, Carla and Manuel Arcedillo. “Machine Translation Evaluation Made Fuzzier: A Study on Post-Editing Productivity and Evaluation Metrics in Commercial Settings.” In *Proceedings of Machine Translation Summit XV*, 2015.
- Piasta, Sbayne B. and Laura M. Justice. “Cohen's d Statistic.” In *Encyclopedia of Research Design*, edited by, 181-85. SAGE Publications, Inc., 2010. <https://doi.org/10.4135/9781412961288.n58>.
- Plitt, M., and F. Masselot. “A Productivity Test of Statistical Machine Translation Post-Editing in A Typical Localisation Context.” *The Prague Bulletin of Mathematical Linguistics* 93 (2010): 7–16. <https://doi.org/10.2478/v10108-010-0010-x>.
- Poibeau, Thierry. “On ‘Human Parity’ and ‘Super Human Performance’ in Machine Translation Evaluation.” In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, 2002.
- Popović, Maja. 2021. “Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output”. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2021.
- Popović, Maja. “Informative Manual Evaluation of Machine Translation Output.” In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.
- . “Error Classification and Analysis for Machine Translation Quality Assessment.” In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Springer, 2018.
- . “Comparing Language Related Issues for NMT and PBMT between German and English.” *The Prague Bulletin of Mathematical Linguistics* 108 (June 2017): 209-220. <https://doi.org/10.1515/pralin-2017-0021>.

- . “chrF: Character n-gram F-score for Automatic MT Evaluation.” In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2015.
- Popović, M., Arcan, M., and Lommel, A. “Potential and Limits of Using Post-edits as Reference Translations for MT Evaluation.” *Baltic J. Modern Computing* 4, no.2 (2016): 218—229.
- Posner, M. I. *Chronometric Explorations of Mind*. Erlbaum, 1978.
- Pym, Anthony. “Redefining Translation Competence in an Electronic Age. In Defence of a Minimalist Approach.” *Meta* 48, no. 4 (2003): 481–497. <https://doi.org/10.7202/008533ar>.
- Pym, Anthony and Ester Torres-Simón. “Is Automation Changing the Translation Profession?” *International Journal of the Sociology of Language* 2021, no. 270 (May 2021). <https://doi.org/10.1515/ijsl-2020-0015>.
- . “Translation Skill-Sets in a Machine-Translation Age.” *Meta* 58, no. 3 (2013): 487–503. <https://doi.org/10.7202/1025047ar>.
- Pyne, R. “Empowerment through Use of the Code of Professional Conduct.” *Br J Nurs*. 3, no. 12 (1994): 631-4. <https://doi.org/10.12968/bjon.1994.3.12.631>.
- Ragni, Valentina and Lucas Nunes Vieira. “What has Changed with Neural Machine Translation? A Critical Review of Human Factors.” *Perspectives* 30, no.1 (2021): 137-158. <https://doi.org/10.1080/0907676X.2021.1889005>.
- Rautio, J. and Koponen, M. *MOLTO Evaluation and Assessment Report*. The MOLTO Consortium, 2013.
- Ray R., DePalma D., Pielmeier H. *The Price-Quality Link*. Common Sense Advisory, 2013.
- Raymond, Eric S. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly and Associates, 1999.
- Resende, Natália. “Investigating Syntactic Priming Cumulative Effects in MT-Human Interaction.” 2021. Open Access: <https://open-research-europe.ec.europa.eu/articles/1-93>. Accessed 25.03.2025.
- Resende, Natália, Benjamin Cowan, and Andy Way. “MT Syntactic Priming Effects on L2 English Speakers.” In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, edited by André Martins, Helena Moniz, Sara

- Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg and Mikel L. Forcada. European Association for Machine Translation, 2020.
- Rico, Celia and Enrique Torrejón. “Skills and Profile of the New Role of the Translator as MT Post-editor.” *Revista Tradumàtica* 10 (2012): 166-78. <https://abacus.universidadeuropea.com/bitstream/handle/11268/657/CR2012.pdf?sequence=1&isAllowed=y>.
- Risku, Hanna. *Translatorische Kompetenz*. Stauffenburg, 1998.
- Robert, Isabelle S., Aline Remael and Jim J.J. Ureel. “Towards a Model of Translation Revision Competence.” *The Interpreter and Translator Trainer* (2016), <http://dx.doi.org/10.1080/1750399X.2016.1198183>.
- Robert, Isabelle, Jim J. J. Ureel and Iris Schrijver. “Translation, Translation Revision and Post-editing Competence Models: Where are We Now?” In *The Human Translator in the 2020s*, edited by Gary Massey, Elsa Huertas-Barros and David Katan. Routledge, 2022. <https://doi.org/10.4324/9781003223344>.
- Robert, Isabelle S., Jim J. J. Ureel, Aline Remael and Ayla Rigouts Terryn. “Conceptualizing Translation Revision Competence: A Pilot Study on the ‘Fairness and Tolerance’ Attitudinal Component.” *Perspectives* 26, no. 1 (2017): 2-23, <http://dx.doi.org/10.1080/0907676X.2017.1330894>.
- Robin, Edina. “The Translator as Reviser.” In *The Modern Translator and Interpreter*, edited by I. Horváth. Eötvös University Press, 2016. <http://www.eltereader.hu/media/2016/04/HorvathTheModernTranslator.pdf>
- Rochard, Michel. “La révision: un acte pédagogique et économique.” In *En bons termes, Actes du colloque international de l'Université de Rennes 2: Spécialités et spécialisations dans la pratique et la formation des traducteurs*, edited by Daniel Gouadec. La Maison du Dictionnaire, 2002.
- Roturier, Johann. “An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical Documentation for French and German Users.” PhD diss, Dublin City University, 2006.
- Rothwell, Andrew, Joss Moorkens, María Fernández-Parra, Joanna Drugan, and Frank Austermuehl, eds. *Translation Tools and Technologies*. Routledge, 2023.
- SAE. “SAE J2450 Translation Quality Metric Task Force.” Accessed May 1, 2025. <https://www.sae.org/standardsdev/j2450p1.htm>.

- Sager, J. C. *Language Engineering and Translation: Consequences of Automation*. Benjamins, 1994.
- Sánchez-Gijón, P., Moorkens, J., & Way, A. “Post-Editing Neural Machine Translation Versus Translation Memory Segments.” *Machine Translation* 33, no. 1-2 (2019): 31–59. <https://doi.org/10.1007/s10590-019-09232-x>.
- Saunders, Danielle. 2022. “Domain Adaptation for Neural Machine Translation.” In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, 2022.
- Scarton, Carolina, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. “Searching for Context: A Study on Document-Level Labels for Translation Quality Estimation.” In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 2015.
- Schäffner Christina, and Beverly Adab. “Developing Translation Competence: Introduction.” In *Developing Translation Competence*, edited by Christina Schäffner and Beverly Adab. Amsterdam: John Benjamins, 2000.
- Schäffner, C. “From ‘Good’ to ‘Functionally Appropriate’: Assessing Translation Quality.” *Current Issues in Language and Society* 4, no.1 (1997): 1–5. <https://doi.org/10.1080/13520529709615476>.
- Schjoldager, A., K. W. Rasmussen, and C. Thomsen. “Précis-writing, Revision and Editing: Piloting the European Master in Translation.” *Meta* 53, no. 4 (2008): 798–813. <https://doi.org/10.7202/019648ar>.
- Scocchera, Giovanna. “The Competent Reviser: A Short-Term Empirical Study on Revision Teaching and Revision Competence Acquisition.” *The Interpreter and Translator Trainer* 14, no. 1 (2019): 19-37. <https://doi.org/10.1080/1750399X.2019.1639245>.
- . “Translation Revision as Rereading: Different Aspects of the Translator’s and Reviser’s Approach to the Revision Process.” *Mémoires du livre / Studies in Book Culture* 9, no. 1 (2017): 1-21. <https://doi.org/10.7202/1043122ar>.
- Secară Alina. “Translation Evaluation: A State of the Art Survey.” In *Proceedings of the eCoLoRe/MeLLANGE workshop* (2005). Accessed May 1, 2025. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e5b3a34db96b2e4ebb4d621bc4f6b8a9735e8f68>.

- Sedon-Strutt, H. "The Revision of Translation Work." *Language International* 2, no. 3 (1990): 28–30.
- Shoshan, O. "Make Neural Machine Translation Better, Faster – a New Way to Measure NMT Quality. Slator. Language Industry Intelligence." Last modified April 19, 2018. <https://slator.com/make-neural-machine-translation-better-faster-a-new-way-to-measure-nmt-quality/>.
- Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. "A Study of Translation Edit Rate with Targeted Human Annotation." In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Association for Machine Translation in the Americas, 2006.
- Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.
- Somers, Harold. *Computers and Translation: A Translator's Guide*. John Benjamins, 2003.
- de Sousa, Sheila C. M., Wilker Aziz, and Lucia Specia. "Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles." In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Association for Computational Linguistics, 2011.
- Specia, Lucia. "Exploiting Objective Annotations for Measuring Translation Post-Editing Effort." In *Proceedings of the Fifteenth Annual Conference of the European Association for Machine Translation*, 2011.
- Stymne, Sara and Lars Ahrenberg. 2012. "On the Practice of Error Analysis for Machine Translation Evaluation." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), 2012.
- Su, Qi, Pavlov, Dmitry, Chow, Jyh Herng, and Baker, Wendell. C. "Internet-Scale Collection of Human-Reviewed Data." In *Proceedings of the 16th International Conference on World Wide Web*, 2007. http://infolab.stanford.edu/~qi/internet_scale_collection_of_human_reviewed_data_ww07.pdf.
- Suojanen T., Koskinen K, Tuominen T. *User-Centred Translation*. Routledge, 2015.
- TAUS. 2010. "Machine Translation Post-editing Guidelines." Accessed May 1, 2025. <https://www.taus.net/resources/reports/mt-post-editing-guidelines>.

- Talman, Aarne, Umut Sulubacak, Raúl Vázquez, Yves Scherrer, Sami Virpioja, Alessandro Raganato, Arvi Hurskainen and Jörg Tiedemann. “*The University of Helsinki Submissions to the WMT19 News Translation Task.*” In *Proceedings of the Conference (WMT19)*. Association for Computational Linguistics, 2019. Doi: 10.18653/v1/W19-5347.
- Temizöz, Özlem. “Translator Post-Editing and Subject-Matter Expert Revision versus Subject- Matter Expert Post-Editing and Translator Revision.” *Journal of Translator Education and Translation Studies* 2, no. 4 (2017): 3-21.
- Terribile, Silvia. “Is Post-Editing Really Faster than Human Translation?” *Translation Spaces* 13, no. 2 (2024): 171-199. <https://doi.org/10.1075/ts.22044.ter>.
- The *Chicago Manual of Style*. 17th ed. The University of Chicago Press, 2024.
- Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. “Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation.” In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, 2018. <https://doi.org/10.18653/v1/W18-6312>.
- Toral, Antonio, Martijn Wieling, and Andy Way. “Post-editing Effort of a Novel with Statistical and Neural Machine Translation.” *Frontiers in Digital Humanities* 5 (2018): 1-11. <https://doi.org/10.3389/fdigh.2018.00009>.
- Tosun, Muharrem, Ayla Akın and Fatih Şimşek. “Courses on Specialized Field in Undergraduate Programs of Translation & Interpretation Departments in Turkey: The Importance of Courses on Specialized Field in the Specialization Process of Translator Candidates.” *Akademik Bakış* 49 (2015): 189–198. <https://dergipark.org.tr/tr/download/article-file/383015>.
- Trofimovich, Pavel and Kim McDonough, eds. *Applying Priming Methods to L2 Learning, Teaching and Research: Insights from Psycholinguistics*. John Benjamins, 2011.
- van Egdom, Gys-Walt and Mark Pluymaekers. 2019. “Why Go the Extra Mile? How Different Degrees of Post-editing Affect Perceptions of Texts, Senders and Products among End Users.” *Journal of Specialised Translation* 31 (2019): 158-176. <https://dspace.library.uu.nl/handle/1874/382993>.
- Valli, Paola. “The TAUS Quality Dashboard.” In *Proceedings of the 37th Conference Translating and the Computer*. AsLing, 2015.
- Vardaro, Jennifer, Moritz Schaeffer, and Silvia Hansen-Schirra. “Translation Quality and Effort Prediction in Professional Machine Translation Post-Editing.” In *Proceedings of the Second MEMENTO workshop on Modelling Parameters of Cognitive Effort in Translation Production*. European Association for Machine Translation, 2019.

- Way, Andy. "Quality Expectations of Machine Translation." In *Translation Quality Assessment: From Principles to Practice*, edited by J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty. Springer, 2018. doi:10.1007/978-3-319-91241-7_6.
- . "Traditional and emerging use-cases for machine translation." In *Proceedings of Translating the Computer 35*, (2013, November 28). <https://aclanthology.org/2013.tc-1.12.pdf>.
- Way, Catherine. "Bridging Professional Practices into Translation Classrooms." In *Proceedings of the Eighth Portsmouth Translation Conference, "The Changing Face of Translation."*, 2009.
- White, John S., Theresa A. O'Connell, and Francis E. O'Mara. "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches." In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 1994.
- Wilks, Yorick. "Corpora and Machine Translation." In *Proceedings of Machine Translation Summit IV*, (1993):137–146. <https://aclanthology.org/1993.mtsummit-1.12/>.
- Wong, Billy T. M. and Chunyu Kit. "Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.
- Wu, Di, Lawrence Jun Zhang & Lan Wei. "Developing Translator Competence: Understanding Trainers' Beliefs and Training Practices." *The Interpreter and Translator Trainer*, 13, no. 3 (2019), 233-254. <https://doi.org/10.1080/1750399X.2019.1656406>.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł, Gouws, S., Kato, Y., Kudo, T., Kazawa, H.,...Dean, J. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." Last revised October 8, 2016. ArXiv manuscript. <https://doi.org/10.48550/arXiv.1609.08144>.
- Yamada, Masaru. "The Impact of Google Neural Machine Translation on Post-Editing by Student Translators." *The Journal of Specialised Translation* 31(2019): 87-106. https://jostrans.soap2.ch/issue31/art_yamada.php.

- . “Can College Students be Post-Editors? An Investigation into Employing Language Learners in Machine Translation plus Post-Editing Settings.” *Machine Translation* 29 (2014): 49-67. <https://doi.org/10.1007/s10590-011-9096-7>.
- Yan, Rui, Yiping Song, Cheng-Te Li, Ming Zhang, Xiaohua Hu. “Opportunities or Risks to Reduce Labor in Crowdsourcing Translation? Characterizing Cost versus Quality via a Pagerank-hits Hybrid Model.” In *Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence*, 2015.
- Yazıcı, Mine. *Yazılı Çeviri Edinci* [Translation competence]. Multilingual, 2007.
- Yücel, Faruk. “Etkili Bir Çeviri Eğitimi.” [An effective translator training.] *Pamukkale University Journal of the Faculty of Education* 22 (2007): 144–155. http://pauegitimdergi.pau.edu.tr/Makaleler/680591722_Faruk%20Y%c3%bcel.pdf.
- Zaidan, Omar F. and Chris Callison-Burch. “Crowdsourcing Translation: Professional Quality from Non-Professionals.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2011.
- Zaretskaya, A. “Raising the TM Threshold in Neural MT Post-Editing: A Case Study on Two Datasets.” In *Proceedings of MT Summit XVII. Volume 2: Translator, Project and User Tracks*. European Association for Machine Translation (EAMT), 2019.
- Zbib, Rabih, Gretchen Markiewicz, Spyros Matsoukas, Richard Schwartz, and John Makhoul. “Systematic Comparison of Professional and Crowdsourced Reference Translations for Machine Translation.” In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2013.
- Zou, Longhui, Michael Carl, Masaru Yamada, Takanori Mizowaki. “Proficiency and External Aides: Impact of Translation Brief and Search Conditions on Post-editing Quality.” In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research)*. Association for Machine Translation in the Americas, 2022.

ANNEXES

ANNEX I: THE TRANSLATION BRIEF

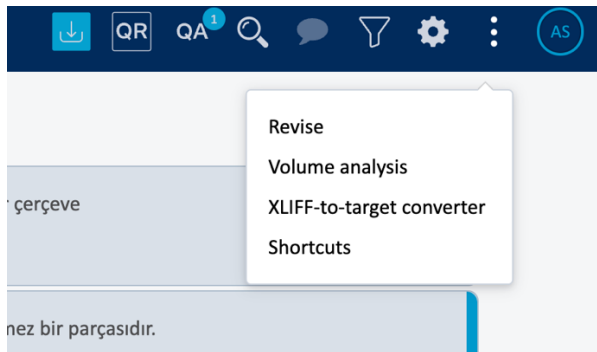
The participants were presented with the instructions document below:

Instructions for the Post-editing and Evaluation Project

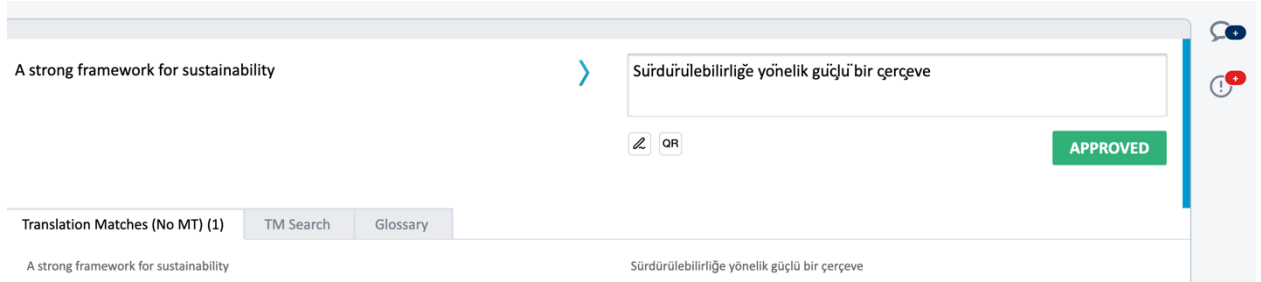
1. After receiving the e-mail for the project, click on the link provided by MateCat.
2. You will see that ST and translation have already been entered in the aligned segments.



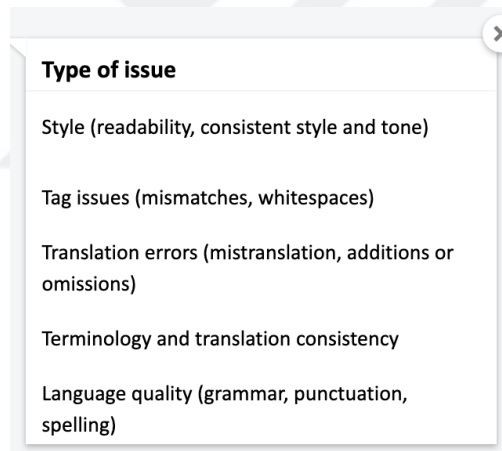
3. Click on the three dots sign on the right-hand side at the top of page and click on Revise.



4. The new page leads to Revision phase where you can both evaluate the translation and post-edit or revise it if you consider it necessary. You are asked to full post-edit the content and evaluate it. That means the final output must be ready to be published once the post-editing is complete. In order to evaluate the translation, simply click on exclamation mark with a red plus sign on it on the segment's right-hand side.



5. The evaluation metric provided by MateCat is in line with TAUS MQM (2024) and is customized as a quick, straightforward and less-detailed alternative. It presents you 5 issue types: style, tag issues, translation errors, terminology and translation consistency and finally, language quality. If you feel the need, you can explore the issue types more in detail on [this website](#). MateCat supports three levels of severity of errors/issues, i.e., neutral, minor and major. You are asked to select one level of error.



6. After post-editing and evaluating the final segment, you do not need to inform the experimenter or click on any save button. The experimenter will be able to follow your progress through MateCat.

Thank you very much for devoting your time to research!

ANNEX II: ST AND TRANSLATIONS USED FOR THE EXPERIMENT

The texts used for PE tasks were 37-segments long. Below are the ST, TT (MT output obtained from Google Translate –NMT – and primed according to Bock 1986 in italics) and TT (Kantan MT output) respectively.

A strong framework for sustainability

Sustainability is an integral part of our business. We are taking our ESG commitment to the next level with our DEGREE sustainability framework. The framework provides a 360-degree approach to our core sustainability values.

By addressing the three aspects of ESG, we are building a better future that helps us to

- → Stay within the planetary boundaries;
- → Foster a culture of trust, empowerment, and growth;
- → Support inclusive economic opportunities;
- → Ensure that our people and businesses remain resilient and

The DEGREE sustainability framework is based on six fields of action that drive sustainability and are dynamic and continuously evolving. We have set clear priorities and ambitions for key ESG issues, which we are driving towards. The DEGREE sustainability framework applies to all Siemens- affiliated companies, excluding Siemens Healthineers (SHS). However, the sustainability approach at SHS follows the same basic principles, which reflects our expectations as its majority shareholder.

Trends and challenges

Global climate change is one of the greatest challenges humanity is facing today, requiring urgent and ambitious action. It is also critical to protect and regenerate our resources and life-support systems like forests, water and biodiversity, for present and future generations. At the same time, we are seeing global poverty and social inequality increase, which in turn jeopardizes human rights and the basis of social coexistence.

Strategy

- – **We combine the real and the digital worlds to scale sustainability impact**
- – **Siemens' business is focused on enabling customers to achieve positive sustainability impact along three customer value propositions: decarbonization & energy efficiency, resource efficiency & circularity, people centricity & societal impact**
- – **The DEGREE sustainability framework defines clear fields of action for our sustainability ambition**

Siemens is a leading technology company with a portfolio designed to drive the digital and sustainable transformation of industry, infrastructure, mobility, and healthcare. We firmly believe that technology is the answer to creating a sustainable future. As key pillars of our strategy, digitalization and sustainability help future-proof our business and that of our customers.

Business units

Our business units (Adhesive Technologies and Consumer Brands) are responsible for adapting our sustainability strategy to their operating needs, as well as providing the resources needed for its implementation. They align their brands and technologies, and the sites involved, to sustainability in line with the specific challenges and priorities of their product portfolio. The research and development departments of our business units work on key technologies and supply the basis for tomorrow's sustainable products.

Siemens offers technologies and solutions to advance growth as societies transition toward a more sustainable future. We enable our customers to accelerate the energy transition, create more resource-efficient factories, smarter buildings, and cleaner transportation, and advance health-care. By combining the real and the digital worlds, we empower our customers to be more competitive and more resilient while simultaneously reducing their environmental impact. Managing data across complex value chains has become an important prerequisite for accelerating this transformation. Siemens' industrial metaverse and digital twin technologies integrate data from physical factories and products across entire lifecycles to optimize production design, increase production efficiency, and reduce CO₂e emissions and the consumption of energy, water, and raw materials. Our products, services, and solutions demonstrate how technology advances sustainability and is helping shape the world we want to live in.

Identifying material topics

To determine and categorize material topics for our business activities and reporting, we engage in a continuous process of recording and evaluating the challenges and opportunities of sustainable development using various tools. We assess the importance and impact of the topics for the company, environment and society, and also for our stakeholders.

Sürdürülebilirlik için güçlü bir çerçeve
Sürdürülebilirlik amaçlı güçlü bir çerçeve

Sürdürülebilirlik işimizin ayrılmaz bir parçasıdır.

DEGREE sürdürülebilirlik çerçevemizle ESG taahhüdümüzü bir üst seviyeye taşıyoruz.
DEGREE sürdürülebilirlik çerçevemiz sayesinde ESG taahhüdümüzü bir üst seviyeye taşımaktayız.

Çerçeve, temel sürdürülebilirlik değerlerimize 360 derecelik bir yaklaşım sunuyor.
Çerçeve ile temel sürdürülebilirlik değerlerimize 360 derecelik bir yaklaşım sunulmaktadır.

ESG'nin üç yönünü ele alarak, bize şu konularda yardımcı olan daha iyi bir gelecek inşa ediyoruz:

ESG'nin üç yönünü ele alarak, bize şu konularda yardımcı olan daha iyi bir gelecek inşa edilecektir:

- → Gezegensel sınırlar içinde kalmak;
- → Güven, güçlendirme ve büyüme kültürünü teşvik etmek;

- → Kapsayıcı ekonomik fırsatları desteklemek;
- → İnsanlarımızın ve işletmelerimizin dirençli kalmasını sağlamak ve *Çalışanlarımızın ve işletmelerimizin dayanıklı kalmasını sağlamak ve*

DEGREE sürdürülebilirlik çerçevesi, sürdürülebilirliği yönlendiren ve dinamik ve sürekli gelişen altı eylem alanına dayanmaktadır.

Temel ESG sorunları için net öncelikler ve hedefler belirledik ve bu hedeflere doğru ilerliyoruz.

Temel ESG sorunları için net öncelikler ve *hedefler belirlenmiştir ve bu hedeflere doğru ilerlenmektedir.*

DEGREE sürdürülebilirlik çerçevesi, Siemens Healthineers (SHS) hariç olmak üzere tüm Siemens'e bağlı şirketler için geçerlidir.

DEGREE sürdürülebilirlik çerçevesi, Siemens Healthineers (SHS) hariç olmak üzere tüm Siemens'e bağlı şirketler için *geçerliliğe sahiptir.*

Ancak SHS'deki sürdürülebilirlik yaklaşımı, çoğunluk hissedarı olarak beklentilerimizi yansıtan aynı temel ilkeleri takip eder.

Ancak SHS'deki sürdürülebilirlik yaklaşımı, çoğunluk hissedarı olarak beklentilerimizi yansıtan aynı temel ilkeleri takip etmektedir.

Trendler ve zorluklar

Küresel iklim değişikliği, insanlığın bugün karşı karşıya olduğu en büyük zorluklardan biridir ve acil ve iddialı eylem gerektirmektedir.

Küresel iklim değişikliği, acil ve iddialı eylem gerektiren, insanlığın bugün karşı karşıya olduğu en büyük zorluklardan biri.

Ayrıca ormanlar, su ve biyolojik çeşitlilik gibi kaynaklarımızı ve yaşam destek sistemlerimizi şimdiki ve gelecek nesiller için korumak ve yenilemek de kritik öneme sahiptir.

Ayrıca şimdiki ve gelecek nesiller için ormanlar, su ve biyolojik çeşitlilik gibi kaynaklarımızı ve yaşam destek sistemlerimizi korumak ve yenilemek de kritik öneme sahip.

Aynı zamanda, küresel yoksulluk ve toplumsal eşitsizliğin arttığını görüyoruz ve bu da insan haklarını ve toplumsal birlikteliğin temelini tehlikeye atıyor.

Aynı zamanda, küresel yoksulluk ve toplumsal eşitsizliğin yükselişte olduğunu görüyoruz ve bu durum, insan haklarını ve toplumsal birlikteliğin temelini tehlikeye atmaktadır.

Strateji

- – Sürdürülebilirlik etkisini ölçeklendirmek için gerçek ve dijital dünyaları birleştiriyoruz – *Sürdürülebilirlik etkisini ölçeklendirmek için gerçek ve dijital dünyalar birleştirilecektir*
- – Siemens'in işi, müşterilerin üç müşteri değer önerisi boyunca olumlu sürdürülebilirlik etkisi elde etmelerini sağlamaya odaklanmıştır: karbonsuzlaştırma ve enerji verimliliği, kaynak verimliliği ve dairesellik, insan merkezlilik ve toplumsal etki

Siemens'te, müşterilerin üç müşteri değer önerisi boyunca olumlu sürdürülebilirlik etkisi elde etmelerini sağlamaya odaklanılmıştır: karbonsuzlaştırma ve enerji verimliliği, kaynak verimliliği ve dairesellik, insan merkezilik ve toplumsal etki

• – DEGREE sürdürülebilirlik çerçevesi, sürdürülebilirlik hedefimiz için net eylem alanları tanımlar

DEGREE sürdürülebilirlik çerçevesi ile , sürdürülebilirlik hedefimiz için net eylem alanları tanımlanmıştır

Siemens, endüstri, altyapı, mobilite ve sağlık hizmetlerinin dijital ve sürdürülebilir dönüşümünü yönlendirmek için tasarlanmış bir portföye sahip lider bir teknoloji şirkettir. *Siemens lider bir teknoloji şirketi olarak endüstri, altyapı, mobilite ve sağlık hizmetlerinin dijital ve sürdürülebilir dönüşümünü yönlendirmek için tasarlanmış bir portföye sahiptir.*

Teknolojinin sürdürülebilir bir gelecek yaratmanın cevabı olduğuna inanıyoruz.

Sürdürülebilir bir gelecek yaratmanın cevabının teknoloji olduğuna inanmaktayız.

Stratejimizin temel direkleri olan dijitalleşme ve sürdürülebilirlik, işimizin ve müşterilerimizin işinin geleceğe hazırlanmasına yardımcı olur.

Stratejimizin temel direkleri olan dijitalleşme ve sürdürülebilirlikle, işimizin ve müşterilerimizin işinin geleceğe hazırlanmasına yardımcı oluyoruz.

İş birimleri

İş birimlerimiz (Yapıştırmacı Teknolojileri ve Tüketici Markaları), sürdürülebilirlik stratejimizi kendi operasyonel ihtiyaçlarına uyarlamaktan ve uygulanması için gereken kaynakları sağlamaktan sorumludur.

Markalarını, teknolojilerini ve ilgili tesisleri, ürün portföylerinin belirli zorlukları ve öncelikleri doğrultusunda sürdürülebilirliğe göre düzenlerler.

İş birimlerimizin araştırma ve geliştirme departmanları, temel teknolojiler üzerinde çalışır ve yarının sürdürülebilir ürünleri için temel oluşturur.

Siemens, toplumların daha sürdürülebilir bir geleceğe doğru geçişi sırasında büyümeyi ilerletmek için teknolojiler ve çözümler sunar.

Siemens, toplumların daha sürdürülebilir bir geleceğe doğru geçişi sırasında büyümeyi ilerletmek için teknolojiler ve çözümler sunmaktadır.

Müşterilerimizin enerji geçişini hızlandırmasını, daha kaynak verimli fabrikalar, daha akıllı binalar ve daha temiz ulaşım yaratmasını ve sağlık hizmetlerini geliştirmesini sağlıyoruz.

Müşterilerimizin enerji geçişini hızlandırması, daha kaynak verimli fabrikalar, daha akıllı binalar ve daha temiz ulaşım yaratması ve sağlık hizmetlerini geliştirmesi sağlanmaktadır.

Gerçek ve dijital dünyaları birleştirerek, müşterilerimizin daha rekabetçi ve daha dayanıklı olmalarını sağlarken aynı zamanda çevresel etkilerini de azaltıyoruz.

Gerçek ve dijital dünyalar birleştirilerek, müşterilerimizin daha rekabetçi ve daha dayanıklı olmalarını sağlarken aynı zamanda çevresel etkileri de azaltılmaktadır.

Karmaşık değer zincirleri boyunca veriyi yönetmek, bu dönüşümü hızlandırmak için önemli bir ön koşul haline geldi.

Karmaşık değer zincirleri boyunca veriyi yönetmek, bu dönüşümü hızlandırmak için önemli bir ön koşul haline gelmiştir.

Siemens'in endüstriyel metaverse ve dijital ikiz teknolojileri, üretim tasarımını optimize etmek, üretim verimliliğini artırmak ve CO2e emisyonlarını ve enerji, su ve hammadde tüketimini azaltmak için fiziksel fabrikalardan ve ürünlerden gelen verileri tüm yaşam döngüleri boyunca entegre eder.

Siemens'in endüstriyel metaverse ve dijital ikiz teknolojileri ile, üretim tasarımı optimize edilmekte, üretim verimliliği artırılmakta ve CO2e emisyonları ve enerji, su ve hammadde tüketimi azaltılmakta ve bunun için fiziksel fabrikalardan ve ürünlerden gelen veriler tüm yaşam döngüleri boyunca entegre edilmektedir.

Ürünlerimiz, hizmetlerimiz ve çözümlerimiz, teknolojinin sürdürülebilirliği nasıl ilerlettiğini ve içinde yaşamak istediğimiz dünyayı şekillendirmeye nasıl yardımcı olduğunu göstermektedir.

Ürünlerimiz, hizmetlerimiz ve çözümlerimiz, teknolojinin sürdürülebilirliği nasıl ilerletebileceğini ve içinde yaşamak istediğimiz dünyayı şekillendirmeye nasıl yardımcı olduğunu göstermektedir.

Önemli konuları belirleme

İş faaliyetlerimiz ve raporlamamız için önemli konuları belirlemek ve kategorilere ayırmak için, çeşitli araçlar kullanarak sürdürülebilir kalkınmanın zorluklarını ve fırsatlarını kaydetme ve değerlendirme konusunda sürekli bir sürece giriyoruz. Konuların şirket, çevre ve toplum için ve ayrıca paydaşlarımız için önemini ve etkisini değerlendiriyoruz.

Below is the Kantan MT output of the ST found above:

Sürdürülebilirliğe yönelik güçlü bir çerçeve

Sürdürülebilirlik, işimizin vazgeçilmez bir parçasıdır. ESG bağlılığımızı, derecesi sürdürülebilirliğimizle bir sonraki seviyeye götürüyoruz. Çerçeve, temel sürdürülebilirlik değerlerine 360 derecelik bir yaklaşım sağlar.

ESG'nin üç unsurunu ele alarak bize yardımcı olmaya yardımcı olan daha iyi bir gelecek oluşturuyoruz

- → Planet sınırları içinde kalın
- → güven, yetkilendirme ve büyümenin bir kültürü.
- → Destek dahil ekonomik fırsatlar:
- → Çalışanlarımızın ve şirketlerin dayanıklı olmalarını ve

Sürdürülebilirlik çerçevesi, sürdürülebilirliğe yol açan ve dinamik ve sürekli gelişmelere sahip altı eylem alanına dayanıyor. Temel ESG sorunları için açık bir öncelik ve ambiyans belirledik. Derecede sürdürülebilirlik çerçevesi, Siemens Healthiners (SHS) hariç olmak üzere tüm Üreticiler bağlı şirketler için geçerlidir. Bununla birlikte, SHS'deki sürdürülebilirlik yaklaşımı, beklentilerimizi çoğaltarak beklentilerimizi yansıtan temel ilkelere uymaktadır.

Eğilimler ve güçlükler

Küresel iklim değişikliği, insanlığın bugün karşılaştığı en büyük güçlüklerden biridir, acil ve hırslı bir hareket gerektirir. Ayrıca, mevcut ve gelecekteki nesiller için ormanlar, su ve biyolojik çeşitlilik gibi kaynakları ve yaşam destek sistemlerimizi korumak ve yeniden oluşturmak da önemlidir. Aynı zamanda, küresel yoksulluk ve sosyal eşitlik artışını görüyoruz, bu da insan haklarını ve sosyal varlığının temelini oluşturmaktadır.

Strateji

- – **Sürdürülebilirlik etkisini ölçeklemek için gerçek ve dijital dünyaları birleştiriyoruz**
- – **Siemens şirketi, müşterilerin üç müşteri değeri prokonumu boyunca olumlu sürdürülebilirlik etkisine ulaşmalarına olanak tanımaktadır: dekarbonizasyon gücü, enerji verimliliği, kaynak verimliliği, insan merkezliliği, sosyal etki**
- – **Sürdürülebilirlik çerçevesinde sürdürülebilirlik tutkumuz için net eylem alanları tanımlar**

Siemens, sektörün, altyapı, mobil kullanım ve sağlık hizmetlerinin dijital ve sürdürülebilir dönüşümünü yönlendirmek için tasarlanmış bir portföye sahip lider bir teknoloji şirkettir. Bu teknolojinin sürdürülebilir bir gelecek yaratmanın cevabının olduğuna inanıyoruz. Stratejimizin temel dayanakları olarak, digitizasyon ve sürdürülebilirlik, işletmemizin ve müşterilerimizin gelişimine yardımcı olur.

İşletme birimleri

İş ünitelerimiz (Yapıştırıcı Teknolojiler ve Tüketici Markaları), sürdürülebilirlik stratejimizi kendi işletme gereksinimlerine uyarlamakla ve uygulanması için gereken kaynakları sağlamaktan sorumludur. Bunlar, marka ve teknolojilerini ve ürün portföylerinin özel güçlükleri ve öncelikleri ile sınırlı bir şekilde mukayabilirliğe sahip sitelere uyum sağlar. İş ünitelerimizin araştırma ve geliştirme departmanları önemli teknolojiler üzerinde çalışır ve yarının sürdürülebilir ürünleri için temel oluşturmaktadır.

Siemens, büyümeyi daha sürdürülebilir bir geleceğe doğru yöneldikçe büyümeyi geliştirmek için teknolojiler ve çözümler sunmaktadır. Müşterilerimizin enerji geçişi hızlandırmasına, daha fazla kaynak verimli fabrikalar, daha akıllı binalar ve daha temiz ulaşım ve daha iyi taşıma olanağı sunmalarını sağlıyoruz. Gerçek ve dijital dünyaları birleştirerek, müşterilerimizin çevre üzerindeki etkisini azaltırken daha rekabetçi ve daha esnek olmalarını sağlamaktayız. Bu dönüşüme hız kazandırmak için karmaşık değer zincirlerinde verileri yönetme, önemli bir ön koşuldur. Siemens endüstriyel metaverse ve dijital ikiz teknolojiler, üretim tasarımını optimize etmek, üretim verimliliğini artırmak ve CO azaltmak için fiziksel fabrikalar ve ürünlerden tüm yaşam döngüleri boyunca verileri entegre eder2e emisyonları ve enerji, su ve ham maddeler tüketimi. Ürünlerimiz, servislerimiz ve çözümlerimiz,

teknolojinin sürdürülebilirliğini nasıl geliştirdiğini ve dünyaya yaşamak istediğimiz dünyayı şekillendirmeye yardımcı olduğunu göstermektedir.

Malzeme konuları tanımlama

İş faaliyetlerimiz ve raporlamalarımız için malzeme konularını belirlemek ve sınıflandırmak için, çeşitli araçları kullanarak sürdürülebilir gelişmenin güçlüklerini ve fırsatları ve fırsatları değerlendirmeye ve değerlendirmeye odaklanıyoruz. Şirket, ortam ve toplum konularının önemli ve etkisini ve ayrıca paydaşlarımız için önemli ölçüde değerlendirilmektedir.



ÖZGEÇMİŞ

Adı, Soyadı	Ayşe Betül		SAYIN
Doğum Yeri ve Yılı			
Bildiği Yabancı Diller	İngilizce (C2)		
ve Düzeyi	Almanca (B2)		
Eğitim Durumu	Başlama - Bitirme Yılı	Kurum Adı	
Lisans	2010	2015	Boğaziçi Üniversitesi
Yüksek Lisans	2015	2019	İstanbul 29 Mayıs Üniversitesi
Doktora	2019	2025	İstanbul 29 Mayıs Üniversitesi
Çalıştığı Kurum/lar	Başlama - Ayrılma Yılı	Çalışılan Kurumun Adı	
1. Biruni Üniversitesi	2023		
2.			
3.			
Üye Olduğu Bilimsel ve Mesleki Kuruluşlar	EST, IATIS, EAMT		
Katıldığı Proje ve Toplantılar	2019 “In Pursuit of ‘Illusion’ in Translation Studies: The Search for Metametonymics in Postcolonial Translation Studies.” II. Uluslararası Rumeli [Dil, Edebiyat, Çeviri] Sempozyumu. 2018 “Çeviribilim Bağlamında Eserlerin Yuvaya Dönüşü Üzerine Betimleyici Bir Çalışma: ‘İllüzyonist’ Çeviri Yaklaşımı ve ‘Metinsiz Geri Çeviriler’.” 25.10.2018. Asoscongress. Yıldız Technical University.		
Yayınlar:	2020. “Modernity Retranslated: Nationalism as a Tool of Culture in a Complex Polysystem.” <i>transLogos</i> 3(1): 70-97. https://doi.org/10.29228/transLogos.22 (MLA- indexed) 2019. “‘Illusion’ in the Postcolonial Context: Translation Studies Still in Need of a Metalanguage or the Understated Lack of Ontology of Science.” <i>transLogos</i> 2(2): 180- 203. doi: https://doi.org/10.29228/transLogos.18 (MLA-indexed)		

	2019. “Çevirinin Yuvaya Dönüşü: Sömürgecilik Sonrası Çeviri Kuramında ‘Metinsiz Geri Çeviriler’”, in <i>Çeviribilimde Araştırmalar</i> (ed. Seda Taş). İstanbul: Hiperyayın, 235-250. (forthcoming in 2025) <i>Dragomanların Rönesansı Diplomatik Mütercimler ve Oryantalizmin Yolları</i> [The Dragoman Renaissance: Diplomatic Interpreters and the Routes of Orientalism, Cornell University Press], ed. Mehmet Kuru (Sabancı). İstanbul: Vakıfbank Kültür Yayınları. 2019. <i>Katolik Korsanlar ve Rum Tüccarlar: Akdeniz’de Bir Ticaret Tarihi</i>. [Translation of Molly Greene’s <i>Catholic Pirates and Greek Merchants: A Maritime History of the Early Mediterranean</i>, Princeton University Press]. İstanbul: Dergâh.
Diğer:	
İletişim (e-posta):	
Tarih İmza Adı Soyadı	6 Mayıs 2025 Ayşe Betül SAYIN

