



MASTERARBEIT | MASTER'S THESIS

Titel | Title

Analysis of translators' productivity gains in the process of
neural machine translation post-editing related to the
translated text's quality

verfasst von | submitted by
Julia Anna Watson

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 070 375 331

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Translation Polnisch Deutsch

Betreut von | Supervisor:

Univ.-Prof. Dragos Ioan Ciobanu PhD

Table of Contents

List of tables	4
List of figures	4
List of abbreviations	6
 1 Introduction	 8
2 Literature overview	10
2.1 Machine Translation	10
2.1.1 Machine Translation approaches	10
2.1.2 Neural Networks and Neural Machine Translation	11
2.2 Post-editing MT	14
2.2.1 Definition of post-editing	14
2.2.2 Teaching post-editing skills	16
2.3 Productivity and quality	18
2.3.1 Productivity – definition and estimation	18
2.3.2 Quality – definition and evaluation	21
3 Methodology	26
4 Results and Discussion	29
4.1 Post-editing productivity and texts' quality	29
4.1.1 Productivity gains and no quality changes	29
4.1.2 Productivity gains and quality losses	34
4.1.3 Productivity and quality gains	35
4.1.4 Productivity losses and quality gains/no changes	38
4.1.5 Chapter summary and conclusion	40
4.2 Relationship between analyzed factors and post-editing productivity and quality	40

4.2.1 Productivity and analyzed factors	40
4.2.2 Quality and analyzed factors	80
5 Conclusion.....	89
5.1 Main findings and implications	89
5.2 Limitations and suggestions for future research	92
References	95
Appendices	104
Appendix I Table with all productivity and quality gains/losses	104
Appendix II Table with all factors influencing productivity and quality gains/losses	106
Appendix III Post-editing and human translation productivity (Jia, et al., 2019a).....	110
Appendix IV Post-editing and human translation productivity (Huang & Wang, 2022)	110
Appendix V Post-editing and human translation productivity (Jia, et al., 2019b).....	111
Appendix VI Post-editing and human translation productivity (Moorkens, et al., 2018)	111
Appendix VII Productivity gains and no quality changes (Stasimioti & Sosoni, 2020)	112
Appendix VIII Productivity gains and no quality changes (Ahmad, et al., 2018).....	113
Appendix IX language codes	113
Appendix X Productivity gains and no quality losses (Cui, et al., 2023)	114
Appendix XI Productivity gains and quality losses (Zhang & Qian, 2023)	115
Appendix XII Productivity and quality gains (Wang, et al., 2021)	115
Appendix XIII Productivity and quality gains (Béchara, et al., 2021)	116
Appendix XIV Productivity and quality gains/losses (Kuzman, et al., 2019)	117
Appendix XV Modified MQM-DQF (Sosoni, et al., 2022).....	117
Appendix XVI Abstract English	118
Appendix XVII Abstract German	119

List of tables

Table 1: Productivity gains and no quality changes	32
Table 2: Productivity gains and quality losses.....	34
Table 3: Productivity and quality gains	37
Table 4: Productivity losses and quality gains.....	38

List of figures

Figure 1: Productivity EN-IT	43
Figure 2: Productivity EN-ZH.....	44
Figure 3: Productivity EN-ES	45
Figure 4: Productivity EN-TR	46
Figure 5: Productivity EN-FI.....	47
Figure 6: Productivity EN-CA.....	47
Figure 7: Productivity: EN-NL.....	48
Figure 8: Productivity EN-PL	48
Figure 9: Productivity EN-HI.....	49
Figure 10: Productivity EN-SL	49
Figure 11: Productivity EN-UK	50
Figure 12: Productivity EN-VI.....	50
Figure 13: Productivity EN-AR.....	51
Figure 14: Productivity EN-EL/EL-EN.....	52
Figure 15: Productivity EN-FR/FR-EN.....	52
Figure 16: Productivity EN-DA	53
Figure 17: Productivity EN-DE.....	53
Figure 18: Productivity EN-PT	53
Figure 19: Productivity EN-SV	54
Figure 20: Productivity for language pairs without English	54
Figure 21: The highest and the lowest productivity for all language pairs	55
Figure 22: Productivity of professional translators	56
Figure 23: Productivity of translation students	57
Figure 24: Productivity of participants with PE experience.....	59

Figure 25: Productivity of participants without PE experience.....	60
Figure 26: Productivity of participants with mixed experience	61
Figure 27: Productivity when PE Google Translate output.....	63
Figure 28: Productivity when PE DeepL output.....	64
Figure 29: Productivity when PE fine-tuned MT output.....	65
Figure 30: Productivity when PE output of remaining MT systems	66
Figure 31: Productivity when PE general texts	70
Figure 32: Productivity when PE literary texts	71
Figure 33: Productivity when PE specialized texts	72
Figure 34: Productivity when PE mixed types of texts	73
Figure 35: Productivity when PE in CAT-tools and using external resources.....	75
Figure 36: Productivity when PE in Translog-II and using external resources	76
Figure 37: Productivity when PE in PET and using external resources	77
Figure 38: Productivity when PE in a Word Document and using external resources	78
Figure 39: Productivity when PE in Word without resources/limited resources.....	79
Figure 40: Quality and language pairs.....	82
Figure 41: Quality and participants	83
Figure 42: Quality and post-editing experience	84
Figure 43: Quality and machine translation systems.....	85
Figure 44: Quality and post-editing different text domains	86
Figure 45: Quality when PE in different experiment settings	87

List of abbreviations

AE	Automatic evaluation
AT	Advertising text
BART	multilingual mBART-50 model
BLEU	Bilingual Evaluation Understudy
COMET	Crosslingual Optimized Metric for Evaluation of Translation
DST	Domain specific text
E	Equal quality
EP	Electronic patents
EX	Experienced
FM	Fine-tuned model
G	Gain
G1.0	Google 1.0
GET	General text
GH	Google high
GL	Google low
GM	Generic NMT model
GNMT	Google Neural Machine Translation
GT	Google Translate
H	Highest gain
HCT	High complexity text
HE	Human evaluation
HT	Human translated
HTME	Major errors for human translation
I	Insignificant
IQG	Insignificant quality gain(s)
IQL	Insignificant quality loss(es)
L&H	Lowest and highest gain
LCT	Low complexity text
LSP	Life science patents
LT	Legal text
M/T	Minutes per source text
MP	Mechanical patents
MQM-DQF	Multidimensional Quality Metrics-Dynamic Quality Framework
MS/S	Milliseconds per source segment
MT	Machine translation
MTE	Machine translation engine
MTH	High-quality machine translation
MTL	Low-quality machine translation
MU	Measure unit
NEX	Not experienced
NM	Not measured
NMT	Neural machine translation
No QE	No quality estimation
NOVEL	Model trained on out-of-domain data and retrained on the novel Practice Makes Perfect

NT	News text
O	outliers
PE	Post-editing
PEME	Major errors for post-edited texts
PEMT	Post-editing machine translation
PT	Poetic text
QA	Quality Assessment
QE	Quality estimation
RQ1	First research question
RQ2	Second research question
S	Significant
S/CH	Seconds per source character
S/S	Seconds per source segment
S/TW	Seconds per target word
S/W	Seconds per source word
SLT	Text about Slovakia
SPOOK	Model trained on out-of-domain data and retrained on the corpus SPOOK (Slovene Translation Corpus)
SQG	Significant quality gain(s)
SQL	Significant quality loss(es)
STH	High-complexity source text
STL	Low-complexity source text
T/P	Time per source page
T/T	Time per source text
T/W	Time per source word
TAUS	Translation Automation User Society
TER	Translation Edit Rate
TLS	Transport layer security
UT	Text about Ukraine
VCT	Venice text
VRT	Verona text
W/D	Source words per day
W/H	Source words per hour
WO	Without outliers

1 Introduction

In the last years, machine translation has become an inseparable part not only of the translation industry, but also everyday life. Daily, great amounts of content are automatically translated, which makes information quickly accessible across countries (Lommel & Pielmeier, 2021). In addition to the substantial volume of texts which are translated by machines daily without being corrected, another essential task, which cannot be overlooked, is post-editing machine translation (PEMT) output.

PEMT involves editing and correcting output that is a result of machine translation (MT), whereas MT is an “automatic translation of text from one natural language to another using a computer application” (ISO:18587, 2017, pp. 1-2). PE of neural machine translation (NMT) output is frequently employed to maintain consistent text quality and enhance productivity compared to human translation (HT). Additionally, it contributes to cost reduction. In theory, it appears to be an excellent approach; putting a text into a machine, subsequently employing a translator to enhance the output and receiving a high-quality product that is both more efficient and cost-effective than producing an entirely new translation from scratch. The reality is that there is no certainty regarding the anticipated outcome. NMT is a relatively recent invention that was introduced in 2016 and is still improving; therefore, quality and productivity of PEMT are still important subjects of research. Questions regarding the actual outcome of PEMT primarily revolve around these two key factors. Firstly, it has been reported that the implementation of PE in the process of NMT brings gains (Sosoni, et al., 2022), however the extent of these increases remains a topic of ongoing discussion. Secondly, there are concerns regarding the quality that varies (Koehn, 2020). Due to the presence of subtle errors that are difficult to detect in MT texts, implementing NMT, even with PE, may not be suitable for every type of text and scenario (Toral, et al., 2018). Finally, another problem occurring is the inconsistent use of metrics and criteria for productivity, as well as for quality, that makes it difficult to quickly and effectively compare results. For instance, time per source page, (Stefaniak, 2020), words per hour (Toral, et al., 2018) or minutes per text (Sosoni, et al., 2022) may be used.

The following thesis analyses translators’ productivity gains through neural machine translation post-editing related to the translated text’s quality, which is included in the first research question: (RQ1): What is the relationship between translators' productivity when post-editing machine translation and the translated text’s quality?

Moreover, the focus is on factors that potentially carry significance regarding quality and productivity. Therefore, the second research question (RQ2) is: Do PEMT productivity and quality differ according to text types, language pairs, types of participants and their PE experience, used NMT systems, and experiment settings?

The data used for the analysis is a text corpus consisting of thirty articles investigating PEMT productivity and quality in the case of NMT. The aim is to determine the relationship not only between PEMT productivity and quality, but also between other important factors. To provide a clear overview of the gathered data, it is organized in tabular form for RQ1, where it is described and compared in detail. For RQ2, each of the factors is set side-by-side with productivity and quality in a chart form, then described accordingly.

This thesis consists of the following chapters: Chapter 2 provides a necessary theoretical basis which includes information on machine translation, post-editing, productivity and quality. Chapter 3 presents the research questions in more detail and discusses the methodology used to select the data analyzed in this thesis. It is followed by Chapter 4 in which the analysis and discussion take place. Finally, Chapter 5 summarizes the main findings and describes suggestions for future research.

2 Literature overview

2.1 Machine Translation

2.1.1 Machine Translation approaches

Machine translation, as well as neural networks have a long story that consists of many boom-and-bust periods (Koehn, 2020, p. 29). Each of these periods had a breakthrough that was followed by high hopes and expectations leading to disappointment and periods in which development would cease (ibid.). However, over the last 20 years, machine translation researchers have been setting much more realistic expectations towards MT.

Machine translation has been present since the 1950s, however its use has evolved over the years and continues to change (Bogucki, 2009). Chronologically, it is possible to differentiate between the following main types of machine translation: rule-based machine translation (RBMT), example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT) (Biel, 2021). NMT is the most current solution that has been commercially used since 2016 (ibid.). It is based on artificial neural networks, deep learning, and large text corpora (Forcada, 2017).

RBMT has different approaches; the first type is a direct translation model that relies on rules to translate from one source language into a target language, while the second type is an interlingua model. In the second one, the content of the source language first must be represented by language-independent codes or symbols to be later translated into the target language. The third model is based on analyzing the source language text, transferring its structure into the target language, and generating the target language text (Hutchins, 1994). For this type of approach, it is necessary to develop sets of rules, syntax, lexis, morphology, and it could pose challenges when translating phrases consisting of metaphors or slang (ibid.).

The second MT type is based on the use of examples. For this purpose, the system is trained with a bilingual corpus, which is used as its main knowledge base, at run time (Somers, 1999). This approach is similar to SMT with the following difference; SMT finds correlations at the word or phrase level depending on aforementioned models, whereas in the case of example-based machine translation, sentences are used as single units that cannot be split (Choudhury & McConnell, 2013, p. 39). For that reason, EBMT requires statistical analysis techniques and needs a carefully aligned corpora that comes with more manual effort (ibid.).

The third main approach to machine translation is SMT. The basis for it are large text corpora that the SMT models are trained on (Koehn, 2020). The corpora can consist of

monolingual texts in the target language, and parallel source language texts and their translations (McEnery & Hardie, 2011, pp. 18-20). Koehn (2020) defines three types of this approach: purely statistical, word-based, and phrase-based.

In the case of SMT, specific sets of rules for each language pair do not need to be defined, as the tools work as they match segments of source texts and translations. SMT, by analyzing the statistical relationships and co-occurrences of words within aligned segments, establishes probabilistic mappings between source and target languages (Choudhury & McConnell, 2013, p. 39). This process is called *alignment* (ibid.). The main problem in the context of these systems is that they are inefficient when translating texts outside of the training corpora (ibid.).

2.1.2 Neural Networks and Neural Machine Translation

NMT is the newest development in the field of machine translation, and it has been rapidly improving in the last years. One of the first successful systems was developed by Google in 2016 (Biel, 2021). Similar to the occurrence and introduction of SMT, this event has expanded the accessibility of NMT to a wider audience, as it enabled rapid translations of information between languages – even quicker than when using previous MT systems (Koehn, 2020).

The word *neural* shows that the inspiration comes from neurons in a brain (ibid.). The neurons in a brain work in the following way: “each neuron receives as input signals from other neurons via its dendrites. If the combined signals are strong enough, the neuron becomes activated and passes a signal via its axon to the various axon terminals that are connected to other dendrites of other neurons” (ibid., p. 30). Artificial neural networks can be described as mathematical inventions that map “an input space into an output space” (Priddy & Keller, 2005, p. 1). The usefulness of artificial neural networks is undeniable. Nevertheless, it is important to remember they are a “man-made invention” that significantly differs from biological neurons whose complexity does not compare with the neural network in the brain (ibid.).

The simplicity of artificial neural networks, in comparison to the natural ones, is visible in the neurons’ connection and activation process. In artificial neurons, activation has its representation in real numbers, whereas “natural neurons have activation in the form of binary spikes at different frequencies” (Koehn, 2020, p. 31). Natural neurons are “connected in orderly architectures” (ibid.), which are complex and organic structures consisting of dendrites, soma, axon, and synapses (Sivanandam & Paulraj, 2009, p. 4). There takes place complex communication in the form of electrical and chemical signals which are received, integrated,

processed and transferred (Nizam, 2024). In the case of artificial neural networks, usually unsupervised training enables the networks to learn their weights (Koehn, 2020, p. 31), meaning it is a simpler system based on the weight sums and activation functions (Nizam, 2024). The output is created by adding up all inputs and passing them through an activation function (ibid.). On the contrary, the evolution of natural neurons happens without a directly given answer or instruction (Koehn, 2020, p. 31).

In an artificial network, neurons behaving in the same way are grouped into specific layers (Sivanandam & Paulraj, 2009, p. 4). All neurons within a layer must either have a full interconnection with each other or none (ibid.). There are three types of layers in a neural network: one input layer, one output layer, and one or more hidden layers. In the input layer, neurons receive external signals and transfer them to the output layer, where signals either from the input or the hidden layer are being received (ibid.). Hidden layers are the ones connected in-between the other two (ibid.). Neural nets are either a single layer network or a multilayer network, whose number can be defined as “the number of layers of weighted interconnection links between various layers” (ibid.). The input layer, however, is not considered when defining the number of layers, since no computation is performed within it (ibid.).

As mentioned, there are two types of neural nets. Perceptrons are the oldest example of the first type, a single layer neural network. Perceptrons were proposed by Rosenblatt (1977) and based on the ideas of McCulloch and Pitts (Koehn, 2020, p. 31). Single layer neural networks are built of an input layer consisting of units that receive signals from the outside (Sivanandam & Paulraj, 2009, p. 4) and one processing layer, which can be described as “a list of neurons, each with the same number of binary inputs and one binary output” (Koehn, 2020, p. 31). A crude learning algorithm proposed by Rosenblatt showed at first the likelihood of perceptrons recognizing different types of patterns (Sabry, 2023). Unfortunately, later, the book *Perceptrons* by Minsky and Papert showed that it was not possible due to the limitations of computer hardware (ibid.). This caused the field of neural networks not to make any significant progress for several years (ibid.).

The research started to develop again only after discovering that neural networks with two or more layers (multilayer perceptrons) possess a greater processing capacity than single layer networks (Sabry, 2023). Multilayer neural networks comprise one or more layers of units, namely hidden layers (Sivanandam & Paulraj, 2009, p. 5). They are made by cascading a group of layers where the input of one layer is provided by the output of the previous layer (ibid.). They are trainable due to their backpropagation algorithm (Koehn, 2020, p. 32). In fact, the algorithm was proposed many times by different researchers, but, because of the neural winter,

not much attention was paid to this subject (ibid.). In 1986, popularizing backpropagation by Rumelhart et al., revived neural network research and gave it new motivation (ibid.).

Additionally, the revival of the research was strengthened by Hornik et al. (1989) and their seminal paper, in which they showed that “any mathematical function could be approximated with neural networks with a sufficient number of layers” (ibid.). After that, neural network research was revived, and networks were trained in different ways, as well as for various purposes and problems (ibid.). At the time, these networks surpassed the quality of any previously used methods (ibid.).

As mentioned, training networks with several layers was a challenge. It applied especially to recurrent neural networks processing sequential data (ibid.). Neural networks were considered potentially useful, however, at the same time, they were thought to be too complex and too difficult to train (ibid.). This was the cause of the second neural winter that started in the late 1990s (ibid.). At that time, the focus was on other machine learning methods that were proposed and broadly used, such as natural language processing (ibid.). Deep learning researchers Geoffrey Hinton, Yann LeCun, and Yoshua Bengio went through a long process before their work was recognized in 2019 with the Turing Award, the highest honor in computer science (Rajarman, 2024, p. 138). The researchers developed better weight initialization methods and activation functions (Koehn, 2020, p. 32). This achievement was a great influence on the development of handwriting recognition task (Rajarman, 2024, p. 137). Later, a big moment came with the graphical processing units (GPUs) and the present wave of neural network research started in 2012 with their submission to the Image classification task ImageNet (Koehn, 2020, p. 32). This convolutional neural network achieved superior results compared to other systems (ibid.).

During the Annual Meeting of the Association for Computational Linguistics in 2015, Chris Manning, a leading natural language processing researcher and professor at Stanford, said “Deep Learning waves have lapped at the shores of computational linguistics for several years now, but 2015 seems like the year when the full force of the tsunami hit the major Natural Language Processing (NLP) conferences” (Manning, 2015 as cited in Koehn, 2020, p. 32).

Nowadays, neural networks are widely used, along with the term “artificial intelligence”. This represents the celebration of the current progress and development (ibid.), and proves AI is no longer just an academic endeavor but a significant part of everyday life (Rajarman, 2024, p. 139). Constant revolutions in artificial intelligence fields are being reported in fields such as healthcare, art, or translation (Koehn, 2020, p. 32) or other specialized fields such tourism (Li & Way, 2021, p. 13). For that reason, also translators are supported by different

innovations. However, their approach differs; some of them celebrate the available technology and try to benefit from it, whereas others have more of a reluctant approach. Regardless of their attitude, it cannot be denied that working with translation technologies has become a part of translators' reality.

2.2 Post-editing MT

Post-editing is becoming one of the most popular tasks in the field of translations due to growing demand for quickly translated texts, as well as widespread use of machine translation. The subsequent subchapters present the definition of PE, explore its various types, and identify the requisite skills for this task, since the thesis investigates quality and productivity in relation to PE of NMT output.

2.2.1 Definition of post-editing

PE is one of the tasks offered in the language industry that can have different forms as described by ISO:18587: 2017. The international standard ISO defines PEMT as editing and correcting output that is a result of MT (ISO:18587, 2017, p. 2). PEMT can further be defined as “a collaboration between a machine and human” with the aim to “provide professional translators with the raw output of the machine translation system and task them to correct it” (Koehn, 2020, p. 21). The PE process is defined by the international standard ISO as it must fulfill certain criteria and consist of specific steps. PE is performed on MT output and its purpose is to check its accuracy and comprehensibility, improve the text, make the text more readable, and correct errors (ISO:18587, 2017, p. 5). The difference between PE and HT is that it involves three texts: the source text, the MT output, and the final target text (ibid.).

There are two types of PE; *full post-editing* that is a “process of post-editing (3.1.4) to obtain a product comparable to a product obtained by human translation” (ibid., 2) and *light post-editing* which is a “process of post-editing (3.1.4) to obtain a merely comprehensible text without any attempt to produce a product comparable to a product obtained by human translation” (ibid.). The use of these two types of post-editing highly “depends on the purpose of the translation output and the client's requirements” (ibid., 5, (Winther Balling, et al., 2014, p. 318). For instance, light postediting can be useful in situations in which quick access to information is needed and in which, therefore, linguistic quality fades into the background. Full

post-editing, on the other hand, would be suitable when the style and other linguistic qualities of the target text play a big role or cannot go unnoticed.

The post-editing process can either be conducted by a human or performed automatically (Winther Balling, et al., 2014, p. 26). Automatic post-editing (APE) “seeks to automatically refine the output of a black-box machine translation (MT) system through human post-edits” (Correia & Martins, 2019, p. 3050). However, this process is often as time consuming as training MT, since the system learns how to correct a machine translated text from a source text, MT target text, and the same MT target text which is human-post-edited (ibid.). For this reason, it is debatable if the implementation of such a system is more advantageous than fine-tuning the existing MT system that translated the source text, using the “newly available post-edited data” (do Carmo, et al., 2021, p. 102). Also, the limited ability to learn of APE systems raises the question whether it is beneficial to pursue the development of MT-systems and APE systems at the same time (Negri, et al., 2018). Nevertheless, APE systems are being explored and modified by researchers. For example, in their article, Correia and Martins (2019) introduce pretrained Bidirectional Encoder Representations from Transformers (BERT) models, that are fine-tuned on the encoder and decoder of an APE system. This approach uses transfer learning, which is proved by the authors to be “an effective and time-efficient strategy for APE” (Correia & Martins, 2019, p. 3050).

The more common type of post-editing, namely human post-editing is defined by ISO: 18587: 2017 in the following way:

“Human post-editing refers to the MT output post-editing process performed by a human post-editor. The post-editing process can take place in an environment where all the text to be processed has been machine translated or in an integrated environment, where MT and TM, along with consistent terminology management, are fully integrated to produce a multi-modal translation work environment. Post-edited MT output can or cannot be fed into a translation memory (automatically or partially), depending on the organization’s processes and its clients’ quality assurance requirements”.

According to the definition, the type of environment needed for post-editing is not strictly defined, as it can mean editing only MT texts or texts that were translated by using MT, TM, and consistent terminology management. Moreover, it is not assumed that the post-edited output must be later fed into a translation memory as it depends on the organization and its client.

2.2.2 Teaching post-editing skills

Given the increasing demand for translation services, which results in a higher number of PE tasks, it has become essential to provide instruction to translators on the PE process. Although it may appear to be a straightforward task, not having experience in PE can cause some confusion. Therefore, the following question emerges: Should PE be taught in translation programs? According to Vasconcellos, developing PE skills is a gradual process (Vasconcellos, 1986, p. 145). As he reports, after having post-edited 100,000 words, translators' level of comfort greatly increases (*ibid.*). Moreover, it has been reported that for some professionals, PE can become not only a more relaxing and enjoyable task than translating from scratch, but it is also less tiring (Stasimioti & Sosoni, 2020; Vasconcellos, 1986; Walker & Federici, 2018), which can be beneficial to the overall job performance. As stated by Somers (1997 p. 201), it is widely believed by many that the skill of PE needs to be "honed" (O'Brien, 2002, p. 99). Being "comfortable" or having some experience with PE is also a skill that many companies desire (O'Brien, 2002; Offer, 2024). One of the factors that might have been keeping translators and universities away from teaching PE is the fear of being replaced by MT. Reluctance to post-edit MT very often results from the fact that rates offered for this type of service are much lower than those paid for pure translation services (Pierce, 2018, p. 214). Another issue seen by translators is in regard to quality. Some translators complain that working with MT lowers their chances to create quality content (Rothwell, et al., 2023, p. 8). For others, post-editing MT-outputs means losing some creativity, since it can be challenging to conceive an alternative phrase or expression after encountering the one proposed by MT, which usually remains close to the source structures (*ibid.*).

On the other hand, many translators who perform PE as a part of their routine work, often report that their tasks are becoming increasingly interesting (O'Brien, 2002, p. 100). As previously noted, contemporary advancements in automated translation technology enable the translation of over a million words within the span of an hour (*ibid.*). Consequently, the time needed for a translation does not play a big role anymore and "the cost-effectiveness of the MT system must be measured largely by the effectiveness of the post-editing process" (Ryan, 1988 as cited in O'Brien, 2002, p. 100). When a human translator post-edits a machine output, their chances of success increase (*ibid.*). There is also a possible change when it comes to translators' approach to MT; it has been confirmed that translators who used to feel threatened by the machine, after working on an MT project, do not have this fear anymore and see it more as an opportunity that they can profit from (Senez, 1998 as cited in O'Brien, 2002, p. 100). However,

it is important to remember that errors made in MT can be difficult to detect and to correct (Rothwell, et al., 2023, p. 8).

Priming effect is, without a doubt, a drawback to PE that warrants attention. Priming can be defined as “the response to a stimulus which is biased by a preceding stimulus” and “an unconscious effect which relies on the implicit memory of the previous stimulus” (Carl & Schaeffer, 2018, p. 55) . Since bilingual minds are based on implicit memory effects (ibid.), translators also belong to this category, therefore, their translation choices are influenced by priming. In this understanding, the bilingual mental lexicon responsible for the priming effect is made of “combinatorial nodes” that connect lexico-semantic and syntactical units in the two languages (ibid.), In the context of PEMT, the output produced by MT is the priming stimulus. MT-post-edited texts usually are more literal than purely translated ones, which could have a detrimental influence on the quality of the translation (Jakobsen & Mesa-Lao, 2017, p. 8).

The research shows that the priming effect is especially dangerous for students who are still unsure about their skills and, as a result, tend to accept more MT suggestions than professional translators. Yamada (2019) in his study reports that due to human-like errors in NMT, students were unable to identify them, which led to more errors in the final translation (Yamada, 2019, p. 87). He stresses that using NMT does not empower students to achieve a translation quality of a professional standard. Additionally, students have shown a lower error correction rate in translation (ibid.). The study revealed that, if human-like translation products of NMT contained errors, students showed difficulties in reaching the required quality standards when post-editing (ibid.).

Despite the disadvantages and doubts regarding PE’s usefulness highlighted above, translators can undeniably profit from it. Incorporating PE into translation courses would yield the following benefits:

- It would contribute to satisfying the growing demand for translation and shorter production times (O’Brien, 2002, p. 100).
- Translators would be equipped with skills specific for PE that are different than translation skills (ibid.). Although a post-editor needs the same basic skillset as a translator (that includes “knowledge about text-type conventions, the ability to deal with style guides and controlled languages, knowledge about contrastive differences, cultural specificities etc.” as well as high proficiency in the source and target language, research skills and world knowledge (Nitzke & Hansen-Schirra, 2021, p. 70), these skills do not guarantee being a successful post-editor (O’Brien, 2002, p. 100). Additionally, post-editors need to be able to assess the

quality of the MT text to determine what corrections are necessary (Nitzke & Hansen-Schirra, 2021, pp. 71-72). Moreover, understanding the limitations of MT helps in handling errors, making it easier to identify, classify, and correct them (ibid.). Finally, consulting competence is essential, since it allows translators to inform their clients, who are often not familiar with PE processes, about possible risks and strategies (ibid.).

- Graduates entering the translation industry would not only be familiar with PE, but also comfortable using it. University training in PE would enable them to quickly and effectively implement MT in their professional life.
- Enhancing translators' understanding of MT and its capabilities could facilitate the adoption of MT technology (O'Brien, 2002, p. 100; Nitzke & Hansen-Schirra, 2021, pp. 70-72).

Educating translators in the field of PE would improve their professional performance and contribute to the expedited advancement of technology in this area. Furthermore, it would boost graduates' professional competence and confidence while expanding their skillset.

2.3 Productivity and quality

Productivity and quality are two factors that play an essential role in translators' work. The quickly growing industry, as well as rapidly increasing demand for making different kinds of products and services available to a wide international audience, influence translators' workflow, which puts pressure on their performance. The goal of this thesis is to analyze translators' productivity gains in the process of neural machine translation post-editing related to the translated text's quality; therefore, it is essential to introduce these two concepts in the context of translation. The first subchapter explains the concept of productivity and presents its definition, as well as its relationship with post-editing. The next subchapter provides an overview of the changing definition of quality, and different perspectives on this concept.

2.3.1 Productivity – definition and estimation

Productivity is “the rate or efficiency of work *especially* in industrial production. It can be expressed as the ratio of output to inputs used in the production process, i.e. output per unit of input” (Cambridge Dictionary, n.d.). It can be influenced by various factors, which means it can

be improved. However, the question of productivity in the context of translation is not quite clear. As TAUS reports, this definition works if there is a list of all variables on the input and output sides; however, a problem may arise if only a limited number of variables is considered (Görög, 2015a). This is frequently the case in the translation industry, where time is viewed as the sole input and words as the only output (ibid.). As a result, productivity is measured by the number of words translated per certain time unit (O'Brien, 2021). As Görög (2015) notices, this approach is simplified, since productivity encompasses far more than merely the time that has been saved. For instance, some studies also considered the number of final edits (Daems, et al., 2017; Jia, et al., 2019a), which can provide a more adequate overview of productivity gains and losses.

It is still not clear what the correlation between machine translation and productivity exactly is, as various statements coming from the industry show. MT engines providers assure that when post-editing its output, translators' productivity increases even up to 350% (Lelner, 2022). There is also research reporting significant increases in translators' productivity when relying on the phrase- and rule-based machine translation engines (Martín & Serra, 2014; Plitt & Masselot, 2010). PE a high-quality MT output usually requires much less effort, therefore, it is expected that the productivity will increase as opposed to using resources of poor quality (Görög, 2015a). In such situations, it could be helpful to take the number of edits per hour into consideration when calculating the productivity score (Daems, et al., 2017; Görög, 2015a).

As previously mentioned, various metrics are used to measure translators' productivity: number of keystrokes, number of words translated per minute (Sosoni, et al., 2022), per hour (Toral, et al., 2018), time spent on translating a source page (Stefaniak, 2020), or simply the overall time spent on translating (Kuzman, et al., 2019). Some of the researchers use multiple factors to calculate the final gains, and, so far, one consistent method, that could generally be used, has not been established. There remains a significant number of variables that would need to be considered in order to accurately measure productivity. Among these variables are:

- Difficulty of the source content (using some measurement independent of language).
- Quality of the source content (based on human assessment by the translator or the reviewer).
- Available resources (also called translation process): whether the translator did or didn't use an MT engine, a translation memory, glossary, external resources, etc.

- Quality of these resources (using fuzzy match and MT confidence information combined with edit distance).
- Number of corrections applied by the reviewer(s).
- Number of errors, weights and penalties applied by the reviewer(s) in the review cycle(s) (Görög, 2015a).

Based on these variables, Nikos Argyropoulos developed a new metric for measuring productivity called TAUS Efficiency Score (*ibid.*). It serves as an alternative to conventional methods of measuring productivity, and can be utilized for evaluating all types of translation: “translation from scratch, translation with translation memory, PEMT or a mix of these three” (*ibid.*). It can be seen as an advantage, since translation jobs increasingly consist of different tasks: one translation job might mean post-editing MT, inserting Translation Memory matches, or fully translating the segments (*ibid.*). The TAUS Efficiency Score can also be seen as a reliable metric, as it is no longer possible to fully separate MT, Translation Memories, and human translation (Görög, 2015b). The TAUS Efficiency Score reflects this phenomenon in the following way: “time is measured for producing (and, if needed, updating) each segment regardless of the segment origin (MT, PE, glossary, scratch, etc.)” (Görög, 2015a). The score is characterized by flexibility; it can be an absolute score based on data from a specific project or a relative score derived from related data in the DQF database (Valli, 2015).

This metric uses two obligatory core variables (words per hour and edit distance) and any additional, optional variables (e.g. difficulty of the source text, experience) (Ruopp, 2015), that are added to the end calculation with the goal of increasing precision and credibility (Görög, 2015a). The score can serve as a metric for evaluating translator efficiency, but it can also be applied to assess the effectiveness of CAT/TMS tools or the performance of MT engines (Valli, 2015).

Measuring the productivity of translators is a complex task, and as a result, various metrics are employed. Researchers may focus on different variables depending on the specifics of a project. Therefore, it is important to mention that it is problematic to compare results of various studies and investigations, in a way that would accurately reflect and provide a comprehensive understanding of productivity gains and losses, in different contexts and situations.

2.3.2 Quality – definition and evaluation

In the context of translation, there are three aspects that really matter: quality, quality and quality (Depraetere, 2011, p. 1). At least this is the prevailing approach within the academic setting. However, upon examining the perspectives of various companies, it becomes evident that speed takes precedence over quality (do Carmo, 2020). From a business standpoint, the time needed for translators to produce the target text is the most relevant factor (Terribile, 2024). The topic of quality comes, however, with different issues. First of all, it is not an easy concept to define. Quality can be looked at from a translator's or an industry's perspective (Depraetere, 2011, p. 2). Additionally, it is important to take into account the perspectives of clients and other actors involved in the process, as well as in any forms of reception, which complicates the concept further.

The concept of quality is closely connected to the expectations set for a translation. For instance, quality in the understanding of the oldest translation norm, namely the sense-for-sense translation, has not been relevant for a long time (Depraetere, 2011, p. 2). According to this norm, creating a word-for-word translation was the main quality indicator (*ibid.*). Another popular approach, born in the XIX century, was called *belles infidèles*. *Les belles infidèles* ("the unfaithful pretty ones") meant "improving" the content of the source text in the translation. That meant, of course, producing a target text with, sometimes more, sometimes less, different content to the source text (Zuber, 1968). In 1969, the *dynamic equivalence* was proposed by Nida (Nida & Taber, 1969). This is when, next to the principle of foreignization, the concept of domestication appeared (Depraetere, 2011, p. 2). Later, Koller, among other researchers, elaborated on this concept (Koller, 1995) and examined various factors, including the communicative context of translation (Toury, 1980) and the social and cultural dimensions (Bassnett & Lefevere, 1998). The role of the context was also later brought to the attention by Reiss and Vermeer (1984), as well as by Nord (1997) and other authors, such as Hatim and Mason, who used pragmatic concepts regarding the specific context of the translation (Depraetere, 2011, p. 2).

As mentioned before, expectations towards translation vary due to its diverse nature and purpose, therefore, expectations towards its quality also differ. All translations come with a certain expectation regarding accuracy and fluency, that can be an indicator for a translation's quality (Koby, et al., 2014). Ramos (2021) however, next to accuracy, names consistency as the main aspects of translation quality, and specifies that these two categories refer to terminology. It is a justifiable approach, since consistency in terminology, and semantic accuracy were

chosen as the first two requirements for quality standards included in ISO17100:2015. According to Koby et al. (2014), “accuracy is a bilingual notion referring to the correspondence between the source and target text, while fluency is a monolingual notion referring to properties of the target text such as grammar, spelling, and cohesion” (Koby, et al., 2014).

There are various definitions of quality. This is a result of diverse views on translation, since translation is a factor defining translational quality (House, 1997, p. 1). For example, ISO defines quality as:

“the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs” (ISO:18587, 2017).

Koby et. al. (2014) propose the following one:

“A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs” (Koby, et al., 2014).

Both definitions refer to fulfilling expectations of the target group, as well as meeting their needs that are simply implied, based on the specifications or stated by the requester and provider. Translation, in the understanding of its broader definition, can mean several activities that include “full translation (that is, segment by segment), summary translation, localization, transcreation, and even gisting (typically a rough translation done by a human or a machine translation system to help people decide whether a professional translation is needed)” (Koby, et al., 2014). The diverse nature of translation is also discussed by House (1997), who makes a remark that it is essential to address the nature of translation before defining what its quality means (House, 1997, p. 1). More specifically, the author discusses the connection between the source text and its translation, as well as how text features affect its reception by various agents. More specifically, the author mentions the relationship between the source text and its translation, between the text features and their perception by agents, as well as the impact that these perceptions have on defining a translation and other textual processes (ibid.). Koby et al. (2014) also point out that there are no absolute specifications that could apply to all translation projects, and therefore, should be made explicit “whenever feasible, relative to the audience and purpose” (Koby, et al., 2014). Since quality is not an easy phenomenon to define, the authors proposed a second, narrowed definition:

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

This definition views translation as text-centric and product-oriented, and does not qualify the activities mentioned by the first definition as a form of translation (Koby, et al., 2014). Moreover, in a case of machine translated texts, it is required that a sufficient level of post-editing is ensured in order to reach minimal levels of accuracy and fluency (ibid.). Many specifications are absolute, even though many of the variables need to be specified (ibid.).

In praxis, quality can be measured in various ways. As mentioned before, defining translational quality depends on translation’s definition, hence the assessment methods differ as well (House, 1997, p. 1). Koehn (2010) explains that there are two main strategies that are used during evaluation: evaluating fluency and adequacy, and ranking of translations (Escartin & Arcedillo, 2015, p. 132). In the case of human evaluation, the judges assess the fluency and adequacy of MT output by assigning a score from one to five on these two criteria (ibid.). Fluency refers to grammaticality, correctness, and idiomatic word choices in the target text (ibid.). Adequacy assesses whether the output segment has the same meaning as the input segment (ibid.). However, there may be instances where the instructions or guidelines are not entirely clear for the reviewers. This situation arises from a disagreement within the industry, as numerous approaches have been created for specific text types, such as literary and advertising texts. (Williams, 2004). The issue is that the approach suitable for these types of text may not correspond with the needs of other text types (ibid.). Moreover, there is a disagreement about which external factors should affect the absolute assessment (ibid.). Finally, all users may have a different idea of quality, therefore it is hard to settle whose definition of quality should be applied and generally used in the evaluation process (ibid.). For that reason, researchers create various quality estimation guidelines.

One of the methods used in the industry, is applying automatic scores such as BLEU, METEOR, or TER, that evaluate MT products by comparing them to reference texts. BLEU (Bilingual Evaluation Understudy) measures the overlap of n-grams between the generated text and reference texts, assigning higher scores to translations that closely match these n-grams (Kolena, n.d.). METEOR considers synonyms and stemming, which enables a more flexible

matching process compared to BLEU (Lavie & Denkowski, 2009). TER (Translation Edit Rate) calculates the minimum number of edits required to change the generated text into the reference text (Machine Translate Foundation, 2023). It focuses on the edit distance as a metric of translation quality (ibid.).

COMET (Crosslingual Optimized Metric for Evaluation of Translation) is an example of neural-network-based scores that are used for evaluating the quality of MT. This automated metric still uses reference translations; however, it differs from previously described systems. COMET, in contrast to BLEU, METEOR, and TER, focuses on semantic similarity by leveraging pre-trained language models to evaluate how well a generated text matches human-like meaning (Lavie, 2020) rather than counting n-gram overlaps. COMET's goal is a deeper understanding of context and semantics, which could provide a more nuanced assessment of translation or produced text's quality (Jiang & Zhang, 2024). Since being launched in 2020, scores like that perform better because they correspond more accurately with human judgement (Mathur, et al., 2020). COMET has been confirmed to be the most accurate system; Sai et al. (2022) confirmed that COMET can be used for low-resource languages, as it corresponds more closely with experts' annotations than metrics that were introduced before (Sai, et al., 2022).

Another way of evaluating quality is using human-based quality assessment or combining it with automatic scores.

For example, The Translation Automation User Society (TAUS) offers various tools and frameworks that can be used for both automatic and human quality evaluation. TAUS is a global organization focused on advancing the translation and localization industry through technology and collaboration (TAUS, 2024). One of its goals is to improve the overall quality and efficiency of localization processes (ibid.).

TAUS offers several complementary approaches that serve different aspects of the evaluation process. Among these, TAUS Quality Assessment (QA) is a systematic framework designed to measure the quality of translation and localization outputs, focusing on criteria that impact user satisfaction and linguistic accuracy (such as accuracy, fluency, terminology consistency, cultural appropriateness, completeness, formatting and style) (Pavlíková, 2022). It provides a template that helps to count, identify, and categorize errors in the translation (ibid.). Based on the template, it is possible to grade the translation's quality, and suggest solutions to identified issues (ibid.). TAUS Quality Evaluation (QE) is a specific component within this framework, focusing on the systematic analysis and scoring of translations to determine their overall quality based on established guidelines and benchmarks (TAUS, 2024). It can involve human evaluators or automated tools to generate scores based on predefined criteria (ibid.).

In contrast, the TAUS Harmonized DQF-MQM Error Typology provides a standardized classification of errors, which makes it easier to evaluate quality across various projects by categorizing errors related to language quality, such as fluency, accuracy, style, variety and terminology (Zhang & Qian, 2023). Lastly, TAUS's Dynamic Quality Evaluation Framework adds flexibility to the QA and QE processes by integrating real-time feedback mechanisms and adaptive metrics that can reflect changing project requirements and user needs (Wang, 2022).

MQM-DQF (Multidimensional Quality Metrics-Dynamic Quality Framework) and TAUS DQF are similar, since they both assess translation quality across multiple dimensions, emphasize critical quality categories such as accuracy, fluency, and terminology, and include them in guidelines, both frameworks support dynamic assessment, and are adaptable for various projects and contexts (Zhang & Qian, 2023).

Quality is challenging to define as is agreeing on one quality evaluation model. While authors may reach different conclusions, and perspectives on quality vary depending on the translation, a few common points emerge: a high-quality translation is accurate and fluent and is expected to meet both the implied and stated needs of its target group.

3 Methodology

This Master's Dissertation analyzes the productivity of translators in the process of PE of NMT and its relation to the quality of the translated text using quantitative descriptive research. The primary objective of this study is to establish the relationship between these two factors, along with other aspects that may influence them.

For the purpose of this research, a text corpus was created. Forty articles were analyzed, and based on the following criteria, thirty of them were selected:

- a) The output of a NMT system in comparison to HT was the object of study.
- b) Information on productivity gains or losses was provided, either in the form numerical or graphical form, and the applied measures were specified.
- c) Productivity was examined in its temporal dimension.
- d) Either quality guidelines were provided, and publishable quality was expected from the participants, or quality assessment was utilized.
- e) Details regarding the language pair, types of participants and their PE experience, type of NMT provider, text type, and experiment setting were included.

The aim was also to collect data that would be as representative as possible. This means that texts concern different language pairs and come from different domains and sources. Ten studies were excluded since they did not comply with all the criteria mentioned above.

The collected data was first organized into two general tables, one for RQ1 and the second for RQ2 (See Appendix I Table with all productivity and quality gains/losses and Appendix II Table with all factors influencing productivity and quality gains/losses). Both tables include the information necessary to answer each research question. In the next step, the data was divided into smaller groups to facilitate the overview of each discussed aspect and to spot relationships between the investigated factors.

To answer RQ1 regarding quality and productivity, the data was organized in a comparative tabular form. The table consists of the following categories: Article, Productivity, Quality and Measure Unit (MU) for productivity and quality. The data was divided into four sections: 4.1.1 Productivity gains and no quality changes 4.1.2 Productivity gains and quality losses, 4.1.3 Productivity and quality gains, 4.1.4 Productivity losses and quality gains/no changes. The results were compared and described to decide whether any productivity and quality gains were

present. It was also analyzed if each article/study discussed both aspects, and if yes, what criteria or metric were used to estimate quality.

One of the goals of this thesis is to present data in a clear and unified manner, therefore, one measure unit was used to demonstrate the findings of different articles. To show productivity gains/losses, percentage values were chosen. However, not all studies have used percentage values. For the purpose of this thesis, percentages were calculated based on the given values; the original numbers were placed in parentheses next to the percentage values. Because of lack of exact numbers in a few articles, approximate gains/losses have been determined based on provided charts, and marked with the $\approx$ symbol. The charts that the estimated numbers refer to can be found in appendices. Some studies have provided more than one result; therefore, acronyms have been placed in parentheses next to percentage values to mark various factors that contributed to different outcomes. The abbreviations used will be presented in each section for ease of reference.

For all categories, various abbreviations were used to describe types of measure units (MU):

- W/H (source words per hour)
- S/S (seconds per source segment)
- M/T (minutes per source text)
- MS/S (milliseconds per source segment)
- T/W (time per source word)
- S/W (seconds per source word)
- S/CH (seconds per source character)
- W/D (source words per day)
- S/TW (seconds per target word)
- T/T (time per source text)

The only type of quality assessment present in the research was human evaluation (HE). It took different forms and followed various criteria; therefore, it will be discussed in detail when needed.

To answer RQ2, the data regarding text types, language pairs, types of participants and their PE experience, used NMT systems, and experiment settings was collected and descriptively presented to examine what kind of dependence occurs. Productivity and quality were compared separately with each of the mentioned factors in subchapters 4.2.1 Productivity and analyzed

factors and 4.2.2, in the form of charts, to ensure clarity. Each subchapter includes seven sections:

1. Productivity/quality and language pairs
2. Productivity/quality and types of participants
3. Productivity/quality and post-editing experience
4. Productivity/quality linked to NMT providers
5. Productivity/quality and text types
6. Productivity/quality and experiment settings
7. Chapter summary and conclusion

Most of articles included in this research show more than one productivity gain/loss. If this is a result of investigating various factors (e. g. types of texts, different groups of participants), the values were included in clustered columns to emphasize the variety. However, if only one factor was examined, but more than one value was available (e. g. a study provided an average gain as well as individual participants scores), the lowest and the highest values were placed above a column. Moreover, if more than two numbers were available, they will be discussed under charts.

Since there are many dimensions of productivity, this research is concentrated only on its temporal aspect. The thesis has the following impact: firstly, it presents the state-of-the-art of PE NMT by comparing actual time benefits and productivity gains and the correlation between productivity boost and quality according to some of the latest research and published reports. Secondly, it sheds more light on the impact of MT use on language service quality. It may also help to set reasonable expectations towards translators and their wages. Furthermore, it provides a useful overview of factors and their influence on PE. Comparing various factors helps to understand the complexity of the PE and translation process and to determine which factors carry significance. Finally yet importantly, it helps to draw attention to the variety of the types of metrics currently used to measure translation productivity and quality.

4 Results and Discussion

4.1 Post-editing productivity and texts' quality

This subchapter addresses the first research question concerning the relationship between post-editing productivity and texts' quality. To enhance clarity in presenting the results, they were divided into six sections. Information in each section is demonstrated in tables composed of the following categories: Article, Productivity, Quality and Measure Unit (MU) for productivity and quality.

Due to the use of various MU in the examined articles, abbreviations were employed. MU utilized in the examined data are as follows:

- source words per hour (W/H)
- seconds per source segment (S/S)
- minutes per source text (M/T)
- milliseconds per source segment (MS/S)
- time per source word (T/W)
- seconds per source word (S/W)
- seconds per source character (S/CH)
- source words per day (W/D)
- seconds per target word (S/TW)
- time per source text (T/T)
- time per source page (T/P)

The only type of quality assessment present in the research was human evaluation (HE). It took different forms and followed various criteria; therefore, it will be discussed in detail when needed.

4.1.1 Productivity gains and no quality changes

Table 1 shows 18 articles reporting productivity gains achieved through post-editing, along with equal quality in post-edited texts when measured against HT texts. In a significant part of the analyzed research, quality has not been reviewed by a third party or evaluated using automatic quality metrics. However, it was possible to include such data in the presented thesis, since, in all cases, the participants were asked to produce texts of “publishable quality” when translating

from scratch and post-editing. Moreover, most of the time, guidelines to assure high-standard quality were provided. That means, in theory, the quality of texts should meet professional standards. However, to avoid confusion when presenting findings, articles that do not measure, and only assume the quality of investigated texts, were marked in the table with the acronym NM (not measured). Research measuring quality was marked with blue color, the highest and the lowest gain were marked respectively with green and purple.

Table 1 includes the following abbreviations:

- Ahmad et al. (2018) demonstrate the highest gain (H) achieved by translators with various levels of translation experience (novice, beginner, medium, expert) for two language pairs.
- Álvarez-Vidal, et al. (2020), Macken, et al. (2020) and Toral, et al. (2018) provide the lowest and the highest gains among translators (L&H).
- Dede (2022) presents results for different qualities of Google Translate; Google high (GH), Google 1.0 (G1.0), Google low (GL). The author determined GT quality by calculating BLUE scores.
- Farrell (2023) investigates PE a text about Ukraine (UT) and Slovakia (SLT).
- Jia et al. (2019a) take into account low (LCT) and high (HCT) complexity texts.
- Jia et al. (2019b) provide two values for domain specific (DST) and general texts (GET).
- Kuzman et al. (2019) discuss results for GT, as well as for a model trained on out-of-domain data and retrained on the corpus SPOOK (SPOOK).
- Murgolo (2022) describes mechanical patents translated with a fine-tuned model (MP, FM) or a generic model (MP, GM), that were also used to translate the following texts: life science patents (LSP FM), (LSP, GM), and transport layer security and electronic patents (TLS/EP, FM), (TLS/EP, GM).
- Sarti et al. (2022) demonstrate gains for various language pairs and two types of models, namely Google Translate (GT) and a multilingual mBART-50 model (BART).
- In Stasimioti & Sosoni's (2020) experiment, participants were experienced (EX) and not experienced (NEX) translators.

- Finally, Terribile (2024) describes data including (O) and excluding outliers (WO)¹.

All language codes used in this table can be found in Appendix IX.

Article	Productivity	MU	Quality	MU
(Ahmad, et al., 2018) ²	96,2% (H, EN-HI novice) 77% (H, EN-HI beginner) 66% (H, EN-HI medium) 161% (H, HI-UR expert) 171% (H, EN-HI expert)	W/D	E	NM
(Álvarez-Vidal, et al., 2020)	53,14% (L&H 33,29-81,01%)	W/H	E	NM
(Dede, 2022)	42,1% (G1.0, mean 29,37) 37% (GH, mean 31,86) 41,5% (GL, mean 29,64)	S/S	E	NM
(Farrell, 2023)	47,99% (UT) 39,5% (SLT)	M/T	E	NM
(Huang & Wang, 2022)	≈50% ³	S/S	E	NM
(Jia, et al., 2019a)	≈55% ⁴ (LCT) ≈72% (HCT)	MS/S	E	NM
(Jia, et al., 2019b)	≈31% (DST) ≈3% ⁵ (GET)	T/W	E	HE
(Kuzman, et al., 2019)	28,7% (GT, PE10.01 vs HT13,0) 4% ⁶ (SPOOK PE52,4 vs HT54,5)	M/T	E	NM
(Läubli, et al., 2019)	59,74% (DE-FR)	W/H	E	HE
(Macken, et al., 2020)	14% (L&H 11-21%)	S/W	E	NM
(Moorkens, et al., 2018)	≈27,25% ⁷	S/CH	E	NM
(Murgolo, 2022)	370% (MP, FM) 324,85% (MP, GM, 889 vs 2,888) 367% (LSP, FM, 683 vs 2,509) 318% (LSP, GM, 683 vs 2,176) 252% (TLS/EP, FM, 720 vs 1,814) 215% (TLS/EP, GM, 720 vs 1,550)	W/H	E	NM

¹ Terribile (2024) describes the process of identifying outliers as follows: “outliers were identified through the mathematical formula applied in SPSS, which uses quartiles to consider the distribution of individual values in the dataset. The first quartile (Q1) is the value under which 25% of cases are found, the third quartile (Q3) is the value under which 75% of cases are situated, and the difference between Q3 and Q1 constitutes the interquartile range (IQR) (Cleff 2014). Outliers are identified through the following equation: Outliers = /< Q1 – (1.5 × IQR) or =/> Q3 + (1.5 × IQR) (Forsyth 2018)” (Terribile, 2024, p. 106).

² Please see Appendix VIII for more detailed data.

³ Please see Appendix IV for a chart used to calculate the approximate value.

⁴ Please see Appendix III for a chart used to calculate the approximate value.

⁵ Please see Appendix V for a chart used to calculate the approximate value.

⁶ Please see Appendix XIV for more detailed data.

⁷ Please see Appendix VI for a chart used to calculate the approximate value.

(Sarti, et al., 2022)	84% (GT, EN-AR) 10% (BART, EN-AR) 119% (GT, EN-NL) 61% (BART, EN-NL) 96% (GT, EN-IT) 95% (BART, EN-IT) 34% (GT, EN-TR) 12% (BART, EN-TR) 34% (GT, EN-TR) 12% (BART, EN-TR) 71% (GT, EN-UK) 14% (BART, EN-UK) 32% (GT, EN-VI) 23% (BART, EN-VI)	S/W	E	NM
(Stasimioti & Sosoni, 2020)	28,3% ⁸ (EX mean 9.86 vs 7.07)	M/T	E	NM
(Stefaniak, 2020)	17%	T/P	E	NM
(Terribile, 2024)	29% (EN-DA, O) 30% (EN-DA, WO) 63% (EN-DE, O) 74% (EN-DE, WO) 67% (EN-ES, O) 67% (EN-ES, WO) 6% (EN-FI, O) 4% (EN-FI, WO) 130% (EN-FR, O) 148% (EN-FR, WO) 89% (EN-IT, O) 90% (EN-IT, WO) 62% (EN-NL, O) 69% (EN-NL, WO) 18% (EN-PL, O) 5% (EN-PT, O) 8% (EN-PT, WO) 5% (FR-EN, O) 86% (FR-EN, WO)	W/H	E	NM
(Toral, et al., 2018)	36% (L&H 1-80%)	W/H	E	NM
(Wang, 2022)	44,7% (2758 vs 223ms)	T/W	E	NM

Table 1: Productivity gains and no quality changes⁹

The table clearly presents that productivity gains vary significantly, while the quality has not changed, according to the researchers' assumptions – in 16 out of 18 cases the quality was not verified. In the given scenario, average research results show the lowest productivity gain of 4% for an English to Finnish text, reported by Terribile (2024). However, taking a closer look reveals that in fact it is 1% (L) according to Toral et al. (2018), although the reported average productivity gain is 36%. Such results may be misleading if the lowest and highest values are

⁸ Please see Appendix VII for a chart used to calculate the approximate value.

⁹ Please see Appendix IX for a full list of the language codes used in this research.

not explicitly disclosed. The highest shown gain amounts to 370% for mechanical patent translated by using a fine-tuned model (MP, FM) as stated by Murgolo (2022). It is crucial to share findings from all participants and scenarios investigated, as they help to illustrate the wide variations in post-editing gains and losses. Relying on an average figure alone does not capture this complexity.

Another point worth noting is the range of gains, not only across all articles, but within individual ones. Namely, Ahmand et al. (2018) report the highest values (HV) ranging from 66% to 171%. Similarly, Terribile's (2024) results show gains between 4% and 148%, when excluding outliers. The same study reports losses in productivity for two language pairs, as shown in Table 4. Sarti (2022) shows results between 10 (for English-Arabic text translated with mBART-50) and 119% (for English-Dutch text translated with Google Translate).

The only study from this category that measured quality and received the exact same value for a PE and HT text, was by Läubli et al. (2019). Läubli (2019) used 1-6 scale (where 1 represents the worst and 6 represents the best) to assess translations and post-edited texts. Experts evaluated these texts based on coherence, cohesion, grammar, style, cultural adequacy, and overall impression. Overall, both texts received 4.50 points. The productivity gain in this case was 59,74%, which is quite a high number that, at least under these circumstances, shows that good quality can certainly come with significant productivity gains when post-editing. It is worth noting, that the same study reported a slight quality gain and a productivity gain under 10% (see Table 3).

The study by Jia et al. (2019b) reported small quality gains and losses for both post-edited and translated from scratch texts, therefore it was still categorized as E. This research shows two lower values that strongly differ from each other, namely $\approx 3\%$ for a general text (GET) and $\approx 31\%$ for a domain specific text (DST). It suggests that PEMT, when applied to domain-specific texts, can offer greater advantages compared to PE in general texts.

Overall, the highest and the lowest values in this section are 1% and 370%. In both cases, the quality was not reviewed, although claimed to be equal. It is important to recognize that translators may have varying interpretations of quality, and there are times when the final product may not meet standards, even with their best efforts. On the other hand, since the quality was not assessed in either case, it cannot be stated that it is to the benefit of post-edited texts due to the high range, that starts with a insignificant gain. The productivity range for texts with measured quality was significantly smaller, between ≈ 3 and 59,7%.

4.1.2 Productivity gains and quality losses

Table 2 shows 6 articles reporting productivity gains achieved through post-editing, as well as quality declines observed in post-edited texts. It is possible to differentiate between two types of quality losses: significant (S) and insignificant (I). The studies included in this section have determined the significance of quality losses by calculating the p value. Each of them is contained in the table in parentheses. The p value is included to prove the significance or insignificance of the losses, since it was used by most of the studies. Only one study, by Pavlíková (2022), did not provide a p value, however, it still described the losses as significant, based on the seriousness of mistakes made in the examined texts. Therefore, the symbol (S) was used, and the number of major errors for human translation (HTME) and post-edited texts (PEME) was included.

The remaining acronyms are as follows:

- Farrell (2018) examined texts about Venice (VCT) and about Verona (VRT).
- Cui, et al. (2023) presented quality losses for a legal text (LT).

As in the previous subchapter, the highest gain was marked with green, and the lowest with purple.

Article	Productivity	MU	Quality	MU
(Ahsan, et al., 2021)	63%	W/H	I (p=0.08)	HE
(Cui, et al., 2023) ¹⁰	53,6% ¹¹ (4.976 vs 2.308mean)	S/TW	I (LT p=.231)	HE
(Farrell, 2018)	5,94% (VCT) 14.94% (VRT)	M/T	I (VCT, p=0.35)	HE
(Hao, et al., 2024)	18%	S/W	<u>S</u> (p < .001)	HE
(Pavlíková, 2022)	29,4% (25.9 vs 36.8min)	T/T	<u>S</u> VS 7,2) PEME:	HE
(Zhang & Qian, 2023)	10,2% ¹² (70.50 vs 63.30mean)	T/T	<u>S</u> (p=0.01)	HE

Table 2: Productivity gains and quality losses

¹⁰ This article can also be found in the table 4.

¹¹ Please see Appendix X for data used to calculate the approximate value.

¹² Please see Appendix XI for data used to calculate the approximate value.

The table includes 6 articles reporting productivity gains and quality losses. Three of them have reported significant losses with productivity gains between 10,2% and 29,4%. The other three studies, showing insignificant quality losses, demonstrate productivity gains ranging from 5.94% to 63%. Notably, Cui et al. (2023) report insignificant losses for a legal text, while also reporting significant quality improvements in other text categories. As a result, their findings are included in Table 3 as well.

Despite human quality evaluators taking part in all the research above, the applied methods and/or criteria took different forms. Ahsan et al. (2021) instructed human evaluators to compare translations and post-edited texts side by side, and to select their preferred option. In Pavlíková's research (2022), reviewers used the TAUS quality assessment template that included thirty-three types of errors covering morphological, lexical, stylistic, syntactic, terminological, formatting, localization, and cultural issues. Farrel (2018) used the following categories: acceptable errors, debatable errors, and mistranslation. Cui et al. (2023) use a 100-mark system with the following criteria: accuracy of meaning, fluency of expression, and reproduction of stylistic characteristics and self-rating, whereas Zhang and Qian (2023) applied the TAUS Harmonized DQF-MQM Error Typology (accuracy, fluency, style, variety). Although the study by Zhang and Qian reports overall worse quality for PE than HT, the quality of the human-translated text was worse in one category, namely for type I accuracy errors (Zhang & Qian, 2023, pp. 103-104). The findings indicate that while HT achieved a superior overall quality score, it is not without errors that could have a significant impact.

Summarizing, the highest productivity gain for this category is 63% and the lowest 5,94%. Three out of six cases indicate significant quality losses through various HE methods, with reported productivity levels falling below 30%. It is worth noting that in all cases of measured quality in this category, productivity gains are not as high as regarding the ones with unmeasured quality (see Table 1).

4.1.3 Productivity and quality gains

Table 3 shows 6 articles reporting productivity and quality gains when post-editing. As the previous tables, Table 3 uses various abbreviations to acknowledge factors discussed in Chapter 4.2.:

- Béchara et al. (2021) use machine translation quality estimation (QE) or no quality estimation (no QE), to investigate its impact on PE.

- Cui et al. (2023) report gains for an advertising text (AT), news text (NT) and poetic text (PT).
- Jia & Sun (2022) published values for high (STH) and low-complexity source texts (STL) translated with the aid of a high (MTH) and low-quality machine translation system (MTL).
- Läubli et al. (2019) provided scores for a text translated from German into Italian (DE-IT).

As previously stated, there were significant (S) and insignificant (I) quality gains. Mostly p. values were available, and therefore, they are presented in the table. The research of Béchara, et al. (2021) did not present very clear scores, hence an estimated percentage score was presented because of the lack of a better measure. Läubli, et al. (2019) and Yang et al. (2023) also do not provide a p. value. However, they present mean numbers, respectively 4,75 vs 4,88 and 84,20 vs 84,58, that were included in parentheses. Results of the articles mentioned in this paragraph were simply marked as a gain (G), since it was problematic to qualify them as significant or insignificant, due to the lack of the p value. The lowest and the highest productivity gain were marked with purple and green.

Article	Productivity	MU	Quality	MU
(Béchara, et al., 2021)	24% (3,8 vs 2,9, no QE) 37% ¹³ (3,8 vs 2,4 s, QE)	S/W	G≈4% (no QE) G≈10% ¹⁴ (QE)	AE
(Cui, et al., 2023)	38,4% (AT, 2.882 vs 4.757) 36,3% (NT, 3.778 vs 2.406) 23,44% (PT, 6.467 vs 4.956)	S/TW	<u>S</u> (AT, p < .001) <u>S</u> (NT, p < .001) <u>S</u> (PT, p < .001)	HE
(Jia & Si, 2022)	32,2% (9.29 vs 7.23s)	S/T	<u>S</u> (p<0.05)	HE
(Jia & Sun, 2022)	76,9% (STH, MTH, 0.13 vs 0.23) 21,4% (STL, MTH, 0.14 vs 0.17) 14,3% (STL, MTH, 0.14 vs 0.16)	W/S	<u>S</u> (STH, MTH p< .0001, STH, MTL p < .05) I (STL, MTH, MTL)	HE
(Läubli, et al., 2019)	9,26% (DE-IT)	TW/H	G (4,75 vs 4,88)	HE
(Wang, et al., 2021)	≈53,9% ¹⁵	W/S	<u>S</u> (p < 0.01)	HE

¹³ Please see Appendix XIII for a chart used to calculate the approximate value.

¹⁴ Please see Appendix XII for a chart used to calculate the approximate value.

(Yang, et al., 2023)	43,4%	(20vs30min)	M/T	G	(84,20vs 84,58 mean)	HE
----------------------	-------	-------------	-----	---	-------------------------	----

Table 3: Productivity and quality gains

As stated before, studies have used different metrics to rate the quality. Béchara et al. (2021) were the only researchers in the gathered corpus who evaluated the quality automatically. They provided FMS scores of the translators' sentences, comparing them to the post-edited reference provided by Autodesk. The researchers also used BLUE, METEOR and word counts to compare the results. Since the scores from all metrics were similar, they were not presented in their research.

The remaining researchers asked human evaluators to rate quality. Cui et al. (2023) applied a 100-mark system with the following criteria: accuracy of meaning, fluency of expression, and reproduction of stylistic characteristics and self-rating. Jia & Si (2022) instructed two professional translators to analyze the results, based on customized error types developed by the Multidimensional Quality Metrics. It included accuracy errors (mistranslation, addition, omission, over-translation and under-translation) and fluency errors (grammar, wrong collocation, logical problem, typing and unintelligible). The other study that Jia conducted with Sun (2022) also calculates the results based on the number of fluency and accuracy errors, that were customized and built on the Multidimensional Quality Metrics. The categories, however, use slightly different subcategories; accuracy (mistranslation, addition, and omission) and fluency (word form, cohesion, word order, punctuation, unintelligible, and spelling). Läubli et al. (2019) used a 1-6 scale (see Table 1). Wang et al. (2021) applied the Taus's Dynamic Quality Evaluation Framework to measure fluency and adequacy. Yang et al. (2023) provided human reviewers with guidelines that were turned into a mean number. The study, however, does not describe the guidelines in detail.

The highest productivity gain, 76,9%, was reported by Jia & Sun (2022), and the measured quality gain was described as significant. The results refer to a high-complexity source text translated with a high-quality machine system. The lowest gain, 9,26%, was presented by Läubli et al. (2019). The quality gain reported for this study was not described as a p value, instead, mean numbers were presented; 4,75 for HT and 4,88 for PE. Overall, six out of eleven quality gains were significant.

Once again, when quality is being measured, the productivity gains are not as high as the ones presented in Table 1, and their range is not as extreme as in the first group (see Table 1).

4.1.4 Productivity losses and quality gains/no changes

Only five out of thirty articles reported productivity losses, together with insignificant (I) quality gains (G), or no changes to quality (E). All of them are presented in Table 4. The lowest number was marked purple.

Abbreviations used in the table below are as follows:

- In the research conducted by Jia and Sun (2022) (STH) means a high-complexity source text, and (MTL) describes a low-quality machine translation model.
- Kuzman et al. (2019) recorded losses for a novel translated by using a model trained on out-of-domain data and retrained on the novel “Practice Makes Perfect” (NOVEL).
- Stasimioti and Sosoni (2020) present a loss for participants with no PE experience (NEX).
- The study of Terribile (2024) demonstrates results for texts including outliers (O) and without outliers (WO).

A full list of language codes used in this research can be found in Appendix IX.

Article	Productivity	MU	Quality	MU
(Jia & Sun, 2022)	-7,7% (STH,PEMTL, 013 vs 0.12)	W/S	I (p > .05)	HE
(Kuzman, et al., 2019)	-16,8% ¹⁶ (NOVEL, PE52,1 vs HT44,4)	M/T	E	NM
(Sosoni, et al., 2022)	-9% (38 vs 41.6min)	M/T	G (21 vs 20 errors)	HE
(Stasimioti & Sosoni, 2020)	-5,7% ¹⁷ (NE, mean 9.86 vs 10.42)	M/T	E	NM
(Terribile, 2024)	-7% (EN-SV, O) -9% (EN-SV, WO) -4 % (EN-PL ¹⁸ , WO)	W/H	E	NM

Table 4: Productivity losses and quality gains

The first glance at the table reveals that productivity losses of PE texts, compared to human translations, did not exceed -17%, as reported by Kuzman et al. (2019).

¹⁶ Please see Appendix XIV for more detailed data.

¹⁷ Please see Appendix VII for a chart used to calculate the approximate value.

¹⁸ Please see Appendix IX for a full list of the language codes used in this research.

The study conducted by Sosoni et al. (2022) was the only one among analyzed articles that described solely productivity losses; in contrast, the other four studies reported productivity gains when taking different factors into account, as detailed in the preceding tables. For instance, the study conducted by Jia and Sun (2022) reported productivity gains for all other investigated categories (see Table 2), and the only loss (-7,7%), visible in Table 4, was recorded for a high complexity source text, translated with a low-quality MT model. The productivity gains from that study ranged between 76,9 and 14,3%, which contrasts with the loss of -7%. A similar outcome is visible in the study by Stasimioti and Sosoni (2020). The researchers report a loss of -5,7% for unexperienced translators (NEX), as well as a productivity gain of 28,3% for experienced translators (EX) (see Table 1). It is important to point out that the study itself focuses more on the average productivity (19,8%). This may result in a misleading interpretation, as a separate analysis of the two groups indicates a considerably different outcome. Terribile (2023) showed a productivity loss of -4% for a text translated from English into Polish. In this case, outliers were not considered (Terribile, 2023). At the same time, she reports an 18% gain for the same text and with the exact same language pair when outliers are taken into account (see Table 1). Moreover, in the same research, Terribile (2023) presents productivity gains between 4% and 148% for ten other language pairs excluding outliers (see Table 1). Only when post-editing a text from English to Swedish, sole productivity losses were reported. Kuzman et al. (2019) report a productivity loss of -16%, when using a model trained on a novel, in contrast to productivity gains (28,7% and 4%) for the same text translated with GT and SPOOK (see Table 1).

According to Table 4, two insignificant quality gains were noted. In the other three remaining cases, the quality was not technically measured, however assumed to be equal by the authors. Quality measures used by Jia and Sun (2022) are described in chapter 4.1.3 *Productivity and quality gains*. Sosoni et al. (2022) applied a modified MQM-DQF error typology¹⁹.

The information provided in the four tables clearly indicates that multiple factors impact productivity, and affect quality gains or losses, even within a single study. Therefore, a few articles had to be placed in two categories. The subsequent chapter will provide a more in-depth discussion of the factors influencing the results.

¹⁹ Please see Appendix XV for more details.

4.1.5 Chapter summary and conclusion

This chapter discussed productivity and its connection to quality in the process of PEMT. The four sections indicate that the majority of research findings highlight productivity improvements without any changes to quality. It is worth noting, however, that most cases do not use any quality evaluation, and the only assurance of the quality is the trust put into the translators' skills. None of the highest productivity gains from the corpus can be found in studies that measure quality. The increase range starts from 1% and reaches 370%, while the lowest value is -16,8%.

Moreover, it became clear that studies generally use various methods to measure productivity and quality. Therefore, bringing all values to the common denominator was necessary; however, establishing this was not always easy due to the unclear nature of some charts. This indicates a need to unify evaluation methods.

4.2 Relationship between analyzed factors and post-editing productivity and quality

This subchapter addresses the second research question regarding the relationship between PE productivity and quality, along with other contributing factors. To facilitate a comprehensive discussion of all relevant factors, each factor was addressed in a separate subsection. Additionally, productivity and quality were analyzed independently in relation to the factors under investigation in the 2RQ.

As previously stated, clarity and transparency are crucial to this research. Therefore, to ensure that information is readily accessible throughout the analysis, detailed descriptions have been included at the beginning of each section. All data was displayed using charts, with percentage values provided for each column.

4.2.1 Productivity and analyzed factors

The following subchapter focuses on productivity and its relationship with factors that may have affected it. These factors include:

1. language pairs;

- English-Polish (EN-PL)
- English-Greek (EN-EL)
- Greek-English (EL-EN)
- English-Danish (EN-DA)
- English-German (EN-DE)
- English-Portuguese (EN-PT)
- English-Swedish (EN-SV)
- German-French (DE-FR)
- Hindi-Urdu (HI-UR)
- German-Italian (DE-IT)
- English-French (EN-FR)
- French-English (FR-EN)
- English-Finnish (EN-FI)
- English-Arabic (EN-AR)
- English-Dutch (EN-NL)
- English-Italian (EN-IT)
- English-Turkish (EN-ER)
- English-Ukrainian (EN-UK)
- English-Vietnamese (EN-VI)
- English-Catalan (EN-CA)
- English-Spanish (EN-SL)
- English-Chinese (EN-ZH)
- Chinese-English (ZH-EN)
- English-Hindi (EN-HI)
- English-Slovenian (EN-SL)
- English-Spanish (EN-ES)
- Spanish-English (ES-EN)

2. types of participants;

- professional translators
- translation students

3. PE experience;

- experienced participants

- unexperienced participants
4. NMT providers/systems;
 - Google Translate
 - DeepL
 - fine-tuned NMT systems
 - other NMT systems (the exact types belonging to this category will be discussed in section 4.2.1.4)
 5. text types;
 - general
 - literary
 - specialized
 - mixed text types (the exact texts belonging to each category will be discussed in section 4.2.1.5)
 6. experiment settings;
 - CAT-tool + access to all types of resources
 - Translog-II + external resources
 - PET
 - working in a Word Document while using other resources
 - translating in a Word document, and having no resources or very limited resources (more detailed description of each of them can be found in section 4.2.1.6).

The only type of post-editing performed in all articles that are part of the research was full PE.

4.2.1.1 Productivity and language pairs

The aim of this subchapter is to gather and review information on productivity and language pairs, and to investigate any potential correlation between these two factors. Since the focus of this chapter is on language pairs and productivity, and an overview of information on multiple productivity gains/losses within an article was presented in Chapter 4.1, the specifics of other factors resulting in more than one percentage value are discussed in subchapters devoted to them.

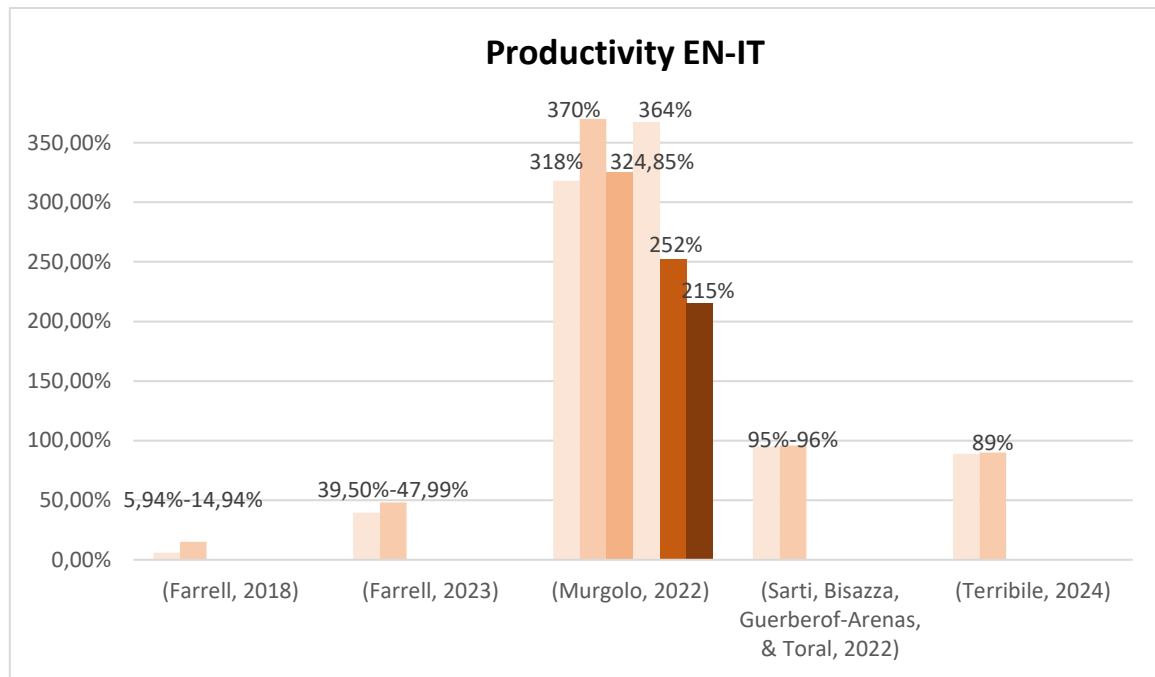


Figure 1: Productivity EN-IT

Among analyzed articles, five investigate EN-IT language pair, and provide scores for various factors:

- Farrell (2018) examined texts about Venice (VCT) and about Verona (VRT).
- Farrell (2023) provides scores a text about Ukraine (UT) and Slovakia (SLT).
- Murgolo (2022) investigated mechanical patents translated with a fine-tuned model (MP, FM) or a generic model (MP, GM), that were also used to translate the following texts: life science patents (LSP FM), (LSP, GM), and transport layer security and electronic patents (TLS/EP, FM), (TLS/EP, GM).
- Sarti et al. (2022) demonstrate gains for various language pairs and two types of models, namely Google Translate (GT) and multilingual mBART-50 model (BART).
- Terribile (2024) describes data including (O) and excluding outliers (WO).

As Figure 1 presents, the range of productivity gains is quite extreme, since the lowest percentage value is just under 6% (Farrell, 2018) and the highest equals 370% (Murgolo, 2022). Based on the figure above, it can be stated that PE the EN-IT language pair can result in unusual productivity gains. Even though only the study by Murgolo (2022) reported unusually high gains, most of research shows gains over 30%. The only exception is shown by the earlier research conducted by Farrel (2018). There could be a few reasons for that: it might be due to the research being the earliest among the analyzed articles, or other influencing factors that will

be addressed in subsequent subchapters. However, it is possible to state that the EN-IT language pair shows potential for great time savings, as it has already exceeded 30%.

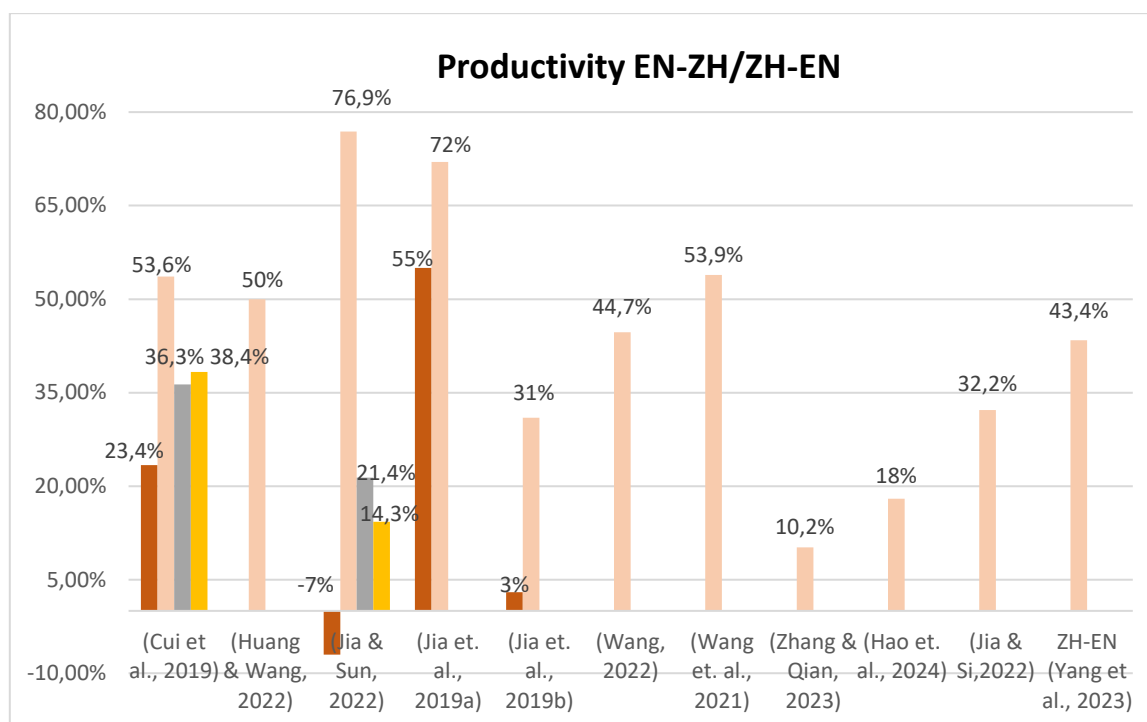


Figure 2: Productivity EN-ZH

Most data for language pairs in the gathered corpus was for EN-ZH. The chart presents ten articles, among which four provide various productivity gains. Different numbers within the same study refer to the following factors:

- Cui, et al. (2023) presented quality losses for a legal text (LT), and gains for an advertising text (AT), news text (NT) and poetic text (PT).
- Jia and Sun (2022) (STH) investigated a high-complexity and low-complexity source text (STL), translated by low-quality and high-quality machine translation models (MTL).
- Jia et al. (2019a) took into account low (LCT) and high (HCT) complexity texts.
- Jia et al. (2019b) provide two values for domain specific (DST) and general texts (GET).

The highest gain is 76,9% (Jia & Sun, 2022). However, the lowest value of -7% was reported by the same research (Jia & Sun, 2022). All remaining studies only present productivity gains, with the variations in values within each individual article being relatively modest. Only one study from the corpus has reported PE productivity for ZH-EN. The productivity amounts to

43,4%, which is close to the average gain for EN-ZH. While the overall values exhibit significant variation, ranging from losses to gains across different levels, it appears that the utilization of post-editing (PE) in this language pair generally facilitates time savings in most scenarios. Nevertheless, due to research reporting losses or quite small gains, translators working with English and Chinese should be cautious. These numbers cannot be overlooked, hence factors causing them will be discussed in the coming chapter.

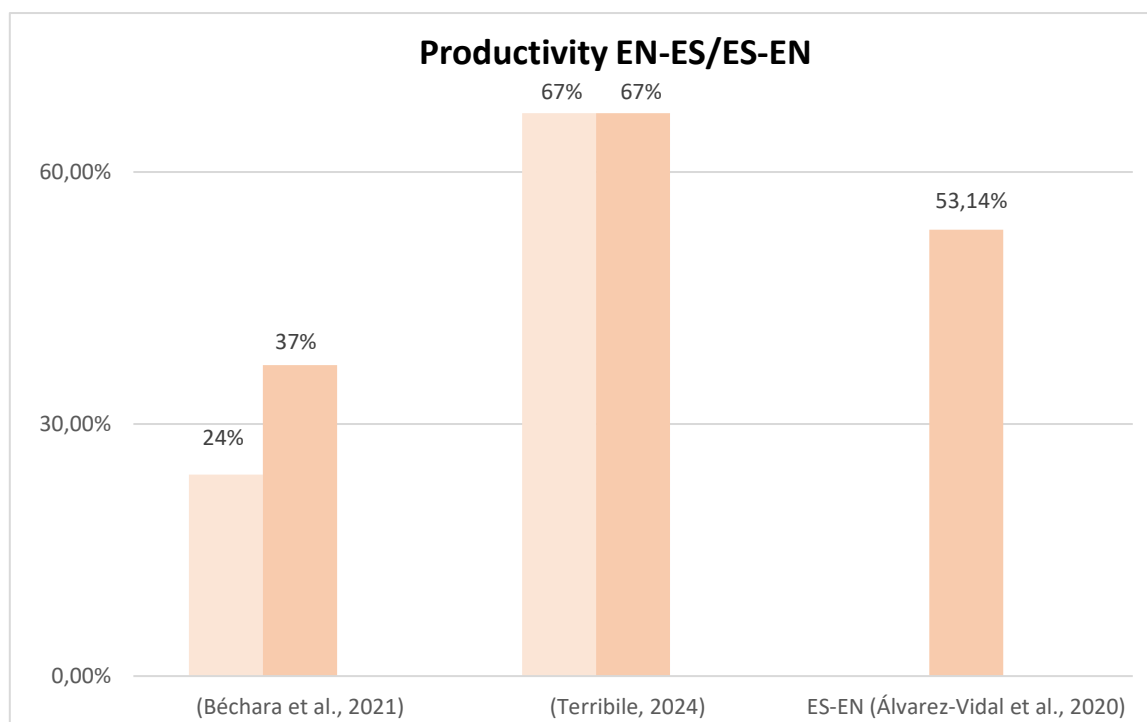


Figure 3: Productivity EN-ES

Figure 3 shows outcomes for PE productivity in language pair EN-ES/ES-EN. All studies have reported pure productivity gains, two of them (EN-ES) for two different scenarios:

- Béchara et al. (2021) applied machine translation quality estimation (QE) or no quality estimation (no QE), to investigate its impact on PE.
- Terribile (2024) describes data including (O) and excluding outliers (WO).

The range of productivity gains visible above is not as big as in the previous studies. It is important to mention that productivity gains for EN-ES (of 67%), reported by Terribile (2024), are similar to the one of 53,14% shown by Alvarez-Vidal et. al. (2020). All studies use the average gain, Alvarez-Vidal et. al. (2020) provide also the lowest and the highest gain; 33.29% and 81.01%. The lower gain is close to the results of Bechara et al. (2021). All the information

above indicates that the discussed language pairs have high time saving potential in both PE directions.

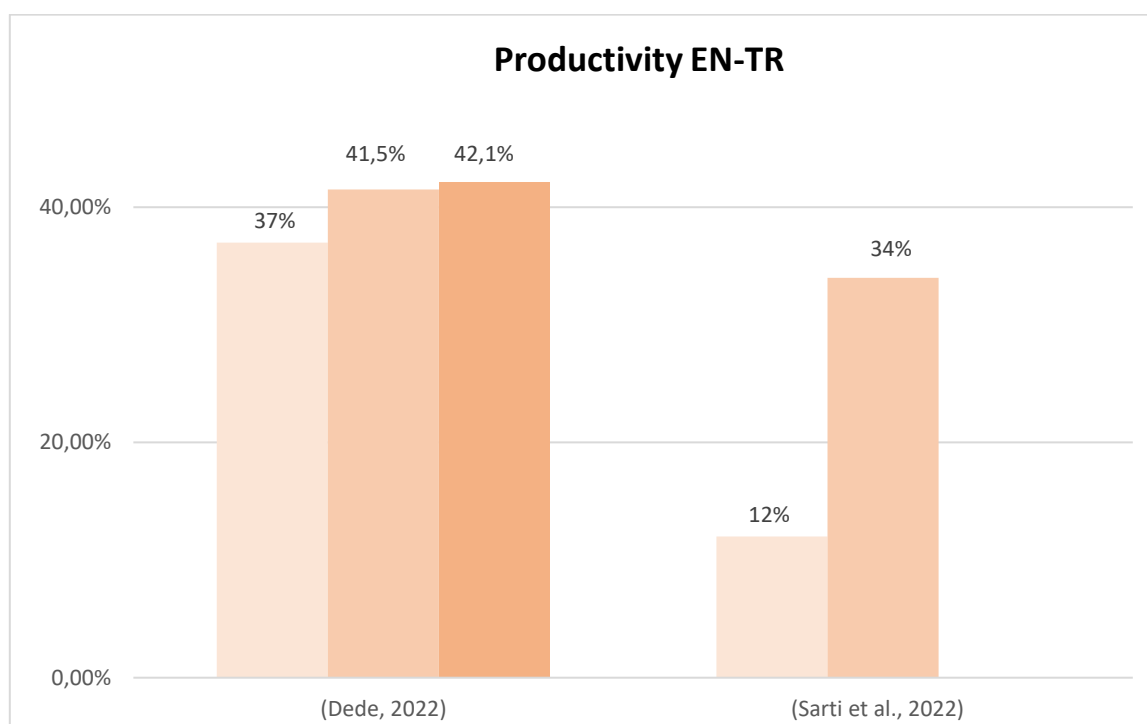


Figure 4: Productivity EN-TR

As displayed in Figure 4, productivity gains for EN-TR were calculated for different MT systems:

- In case of Dede (2022), various qualities of GT were discussed: Google high (GH), Google 1.0 (G1.0), Google low (GL). The author determined GT quality by calculating BLUE scores.
- Sarti et al. (2022) used Google Translate (GT) and a multilingual mBART-50 model (BART).

The lowest gain for post-edited EN-TR texts is 12% in study by Sarti et al. (2022), and the highest one 42,1%. It is, however, worth mentioning that Dede (2022) focuses on the average gain of 40,2%, although numbers for specific machine translation systems are provided. The difference between the lowest and the highest percentage value does not seem big, nevertheless, it could have significance for translators when setting their wages.

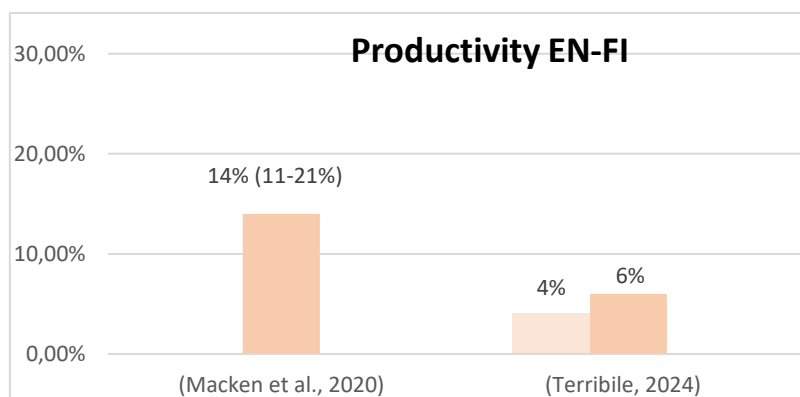


Figure 5: Productivity EN-FI

Productivity gains for EN-FI post-edited texts are not as high as for other language pairs. The highest gain equals 21% and the lowest 4% which comes down to 17% variation. Even though Macken et al. (2020) present the lowest and the highest number, they base their results on the average value of 14% which then is more comparable with the gains presented by Terribile (2024), since the research done by her also describes average gains for translation with and without outliers. The researchers also note that when translators are provided with perfect MT suggestions, that do not require any edits, they can work 66% faster compared with translating from scratch (Macken, et al., 2020).

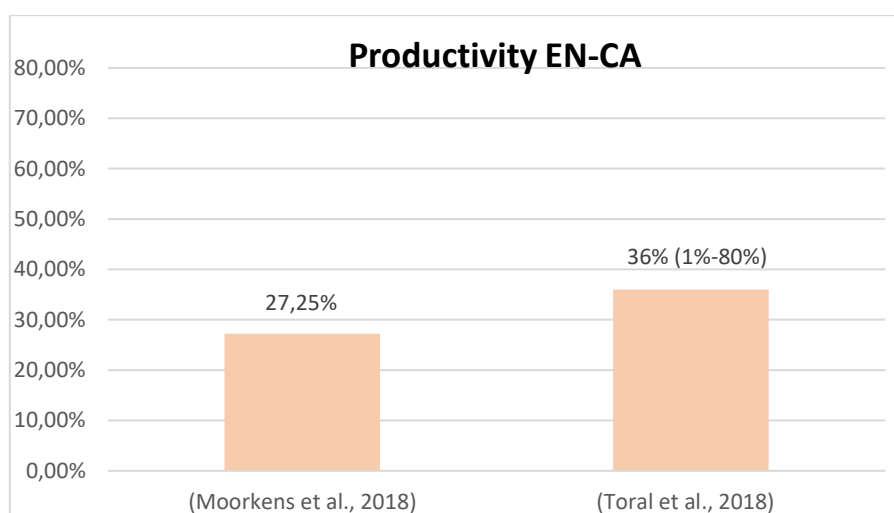


Figure 6: Productivity EN-CA

Toral et al. (2018) present an average gain of 36% which is close to 27,25% reported by Moorkens et al. (2018). The analysis of individual scores by Toral et al. (2018) indicates actual improvements ranging from 1% to 80%. This suggests that the higher gains observed for this language pair in the study by Moorkens et al. (2018), which reported average scores, are likely as well.

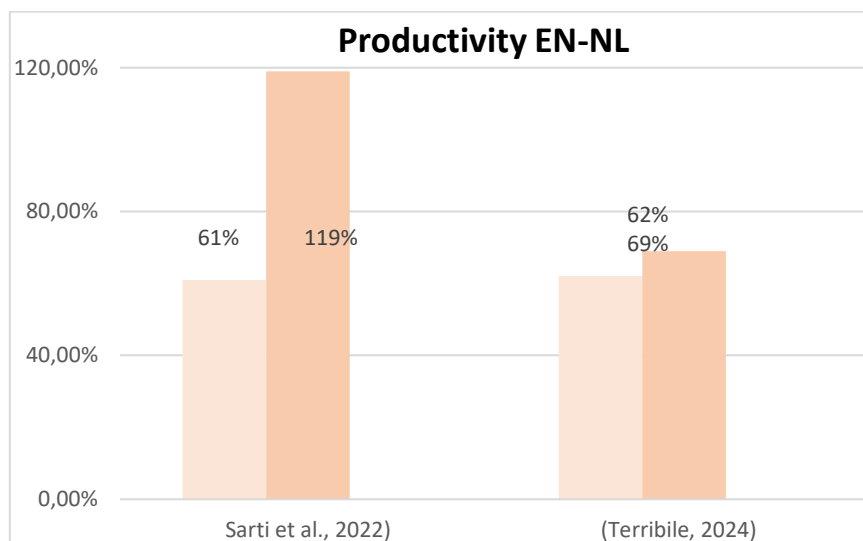


Figure 7: Productivity: EN-NL

Productivity gains in the case of MTPE for EN-NL texts are high; three out of four numbers are close to each other. The highest score was the outcome of using Google Translate (the second score reported by Sarti et al. (2022) was for the mBART system). Scores calculated for text with and without outliers are very similar (Terribile, 2024). The gains vary depending on the system in use, but regardless of the MT provider selected, employing NMT for this language pair appears to yield substantial profits.

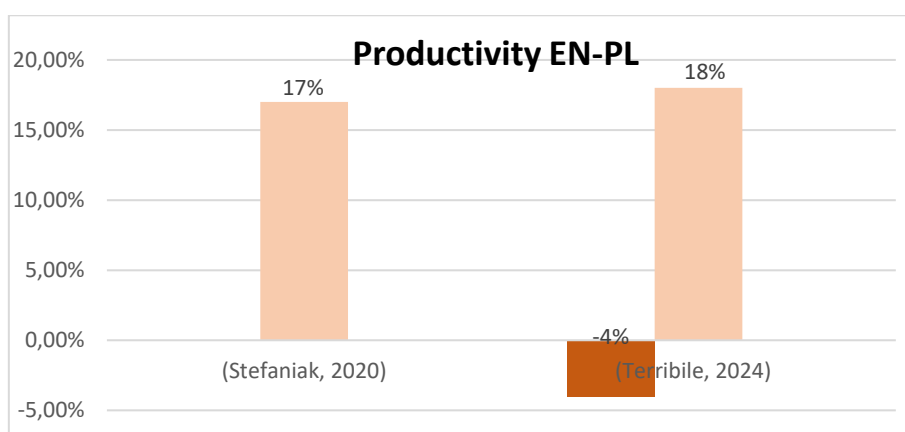


Figure 8: Productivity EN-PL

Productivity increases described in both articles are exceptionally close -they differ by 1%. Despite that, Terribile recorded a productivity loss of -4%, however it does not strike as much as in cases of other studies, since the gains are not very high. A reason for that could be the complexity of Polish, and the lack of similarities between EN and PL, as well as the fact that the EN-PL language combination is not highly resourced in comparison to other pairs.

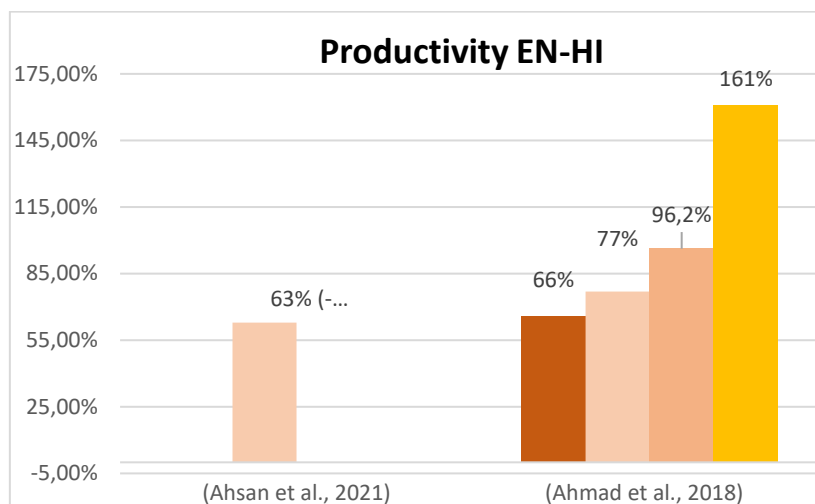


Figure 9: Productivity EN-HI

Both studies provide more than one productivity value. Ahsan et al. (2021) focus on the average gain of 63%. Taking a closer look at the data reveals, however, that productivity differs from -2% to 150.2%. Ahmad et al. (2018) show pure gains for participants with different levels of translation expertise, namely novice, beginner, medium and advance translators. The numbers reported are as follows: 96.2%, 77%, 66% and 161%. On average both studies report significant gains that are even higher for certain participants. Overall, PE when translating from EN to HI can be described as profitable apart from one case.

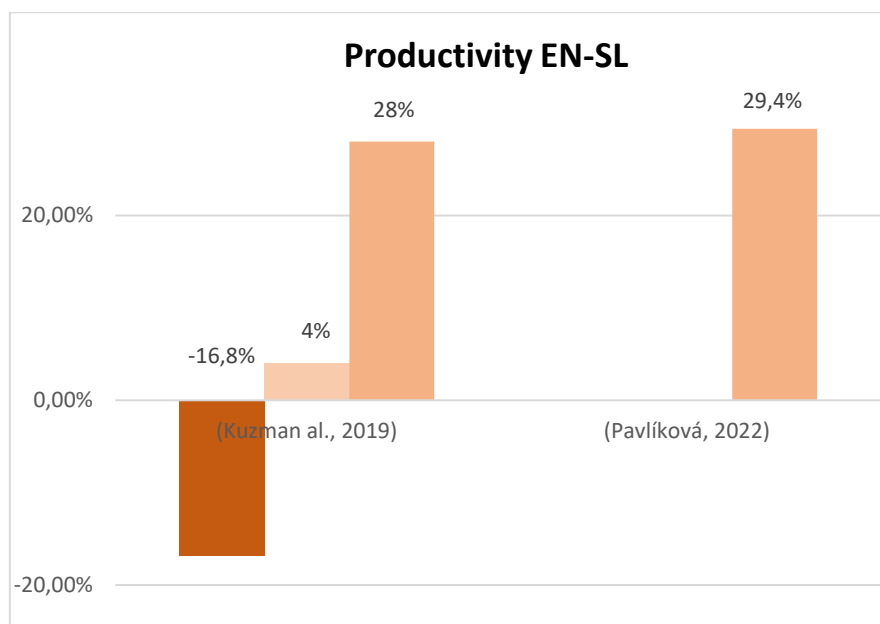


Figure 10: Productivity EN-SL

Figure 10 displays productivity for EN to SL translations. Kuzman et al. (2019) utilized three various NMT systems; GNMT, a model trained on out-of-domain data and retrained on the

novel “Practice Makes Perfect”, and a model trained on out-of-domain data and retrained on the Slovene Translation Corpus. The highest score reported in both studies was for GNMT (GT was also utilized by Pavlíková (2022)), and the lowest for the model NOVEL. It seems that gains achieved with GNMT for this language pair may be stable, regardless of the text type (Pavlíková (2022) investigated specialized texts). However, further research is necessary to validate this conclusion.

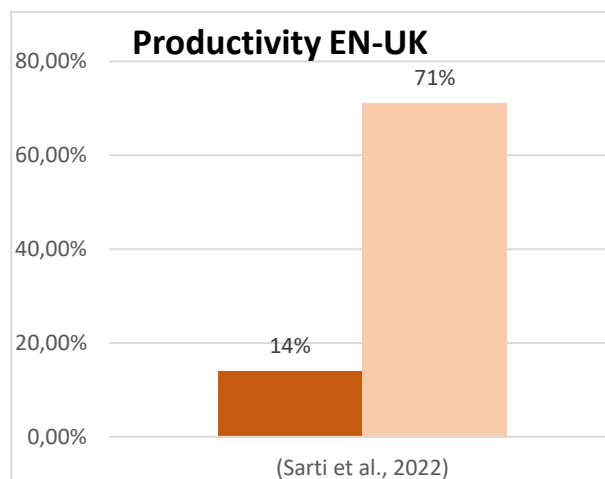


Figure 11: Productivity EN-UK

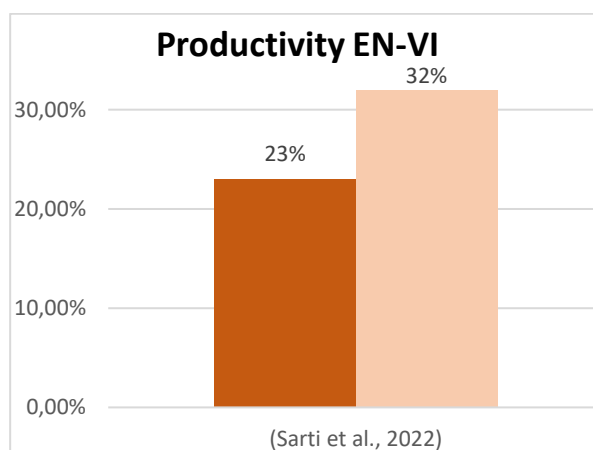


Figure 12: Productivity EN-VI

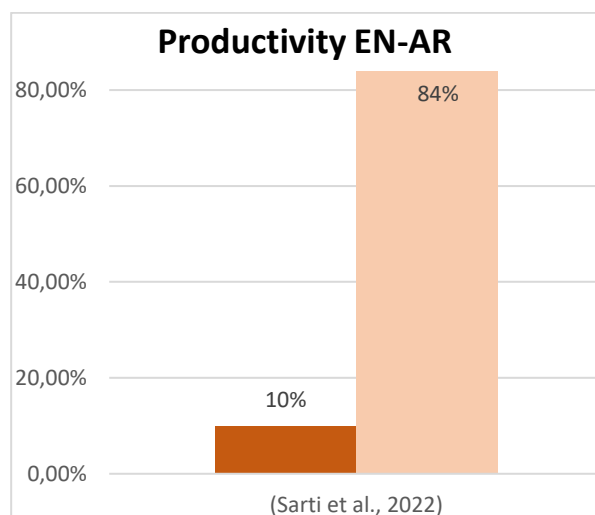


Figure 13: Productivity EN-AR

Figures 11, 12 and 13 present productivity gains for the following language pairs: EN-UK, EN-VI and EN-AR. Only one research, conducted by Sarti et al. (2022), describes PE texts in all these language pairs. Therefore, this thesis cannot compare the results with other sources, and the language combinations mentioned above will be discussed together in the next paragraph.

GT and mBART were the NMT systems investigated in this research. Two gains were reported for each language pair: the ones for EN-AR and EN-UK are quite different, 10%-84% and 14%-71% respectively. All lower scores were reported for mBART, and all the higher ones for GT. Results for EN-UK and EN-VI vary strongly when comparing these two systems, while EN-VI do not differ so drastically (23% and 32%). All language pairs mentioned above may bring productivity gains if the right machine translation system is used. As mentioned, it is to be remembered that more research is needed to support or contradict this conclusion.

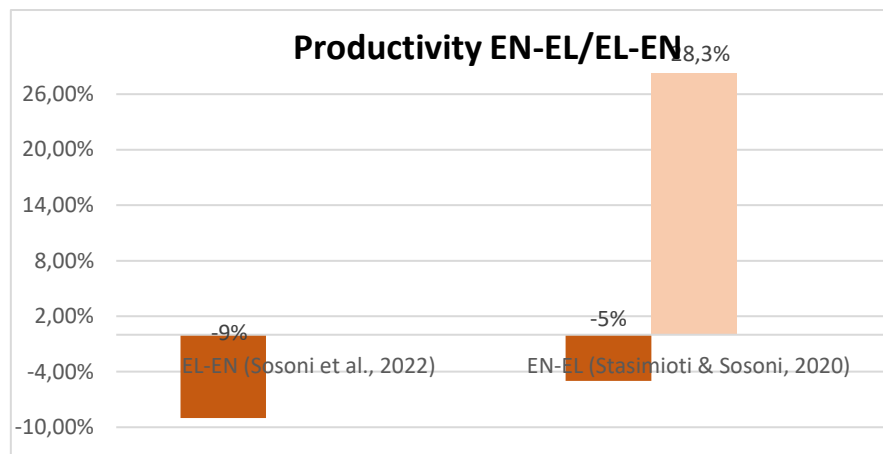


Figure 14: Productivity EN-EL/EL-EN

Figure 14 shows the correlation between productivity when post-editing the language combination EN-EL and EL-EN. Overall, the data shows that it is not easy to gain productivity in the language pair visible in Figure 14, unless there is significant experience in PE, as the highest value of 28,3% was observed among participants with PE experience. In contrast, those with no experience showed a loss of 5%, while translators who had only undergone a brief training session prior to joining the experiment experienced a loss of 9%. Another reason for speed losses in research conducted by Sosoni et al. (2022) could be a limited number of participants – only two translators took part in it.

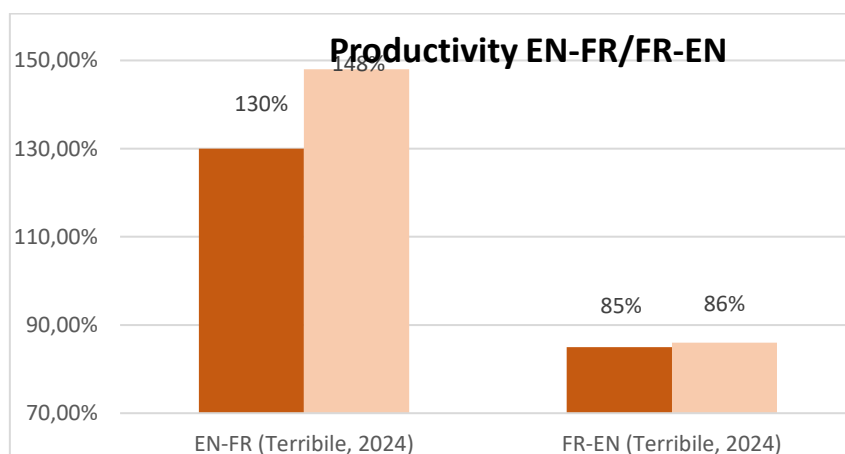


Figure 15: Productivity EN-FR/FR-EN

Terribile (2024) is the only one among chosen researchers reporting productivity for EN-FR and FR-EN language combination. At the first glance, it is evident that, although PE in both directions brings big gains, the gains are much higher for the EN to FR direction even within the same research. Moreover, both categories (higher scores with outliers and lower without outliers) for each translation direction show similar numbers.

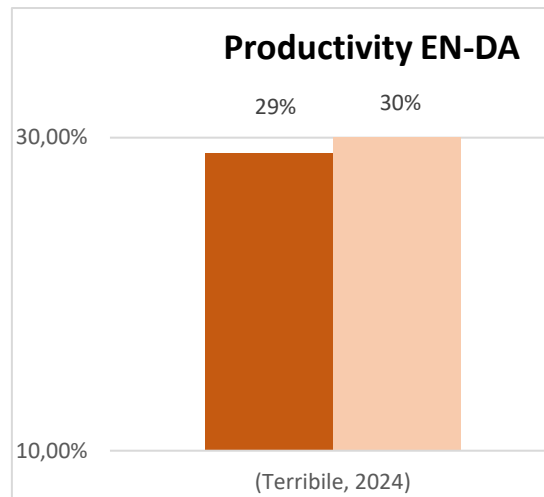


Figure 16: Productivity EN-DA

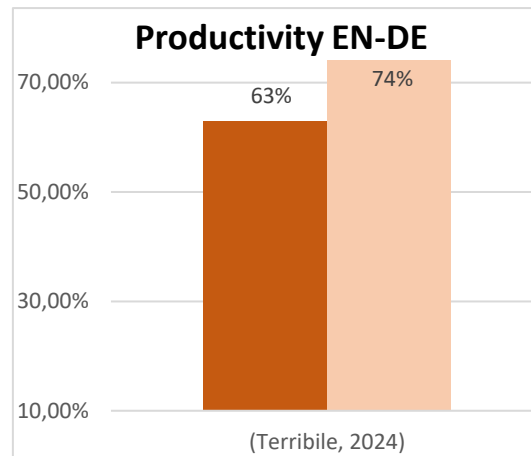


Figure 17: Productivity EN-DE

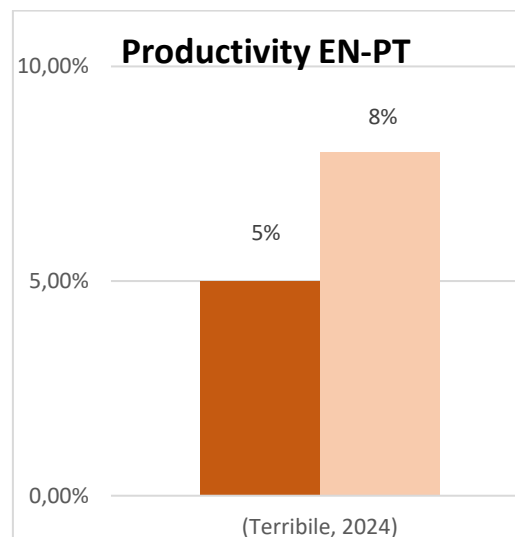


Figure 18: Productivity EN-PT

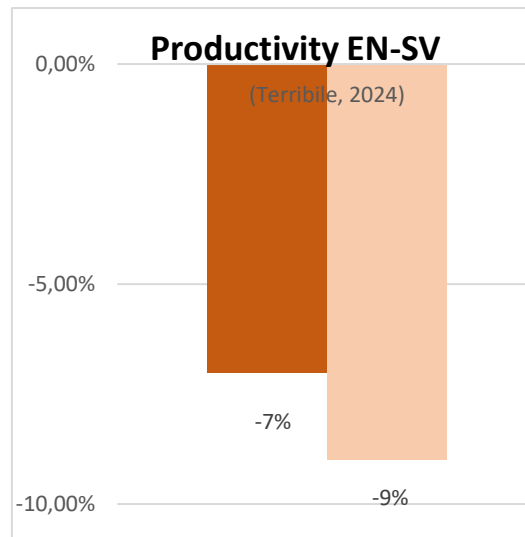


Figure 19: Productivity EN-SV

Study conducted by Terribile (2024) was the only one to investigate EN-DA, EN-DE, EN-PT and EN-SV language pairs. In the gathered corpus, there are no other studies that could be used for comparison; therefore, these language combinations will be discussed together in one section. Figure 19 presents interesting results, as it is the only one to report productivity losses in both categories in the discussed study. Moreover, the loss of 9% was the lowest among all other language pairs. The gains for EN-DA and EN-DE would be sufficient, since they exceed the threshold of 30%. Productivity increase for EN-PT is relatively low and insignificant.

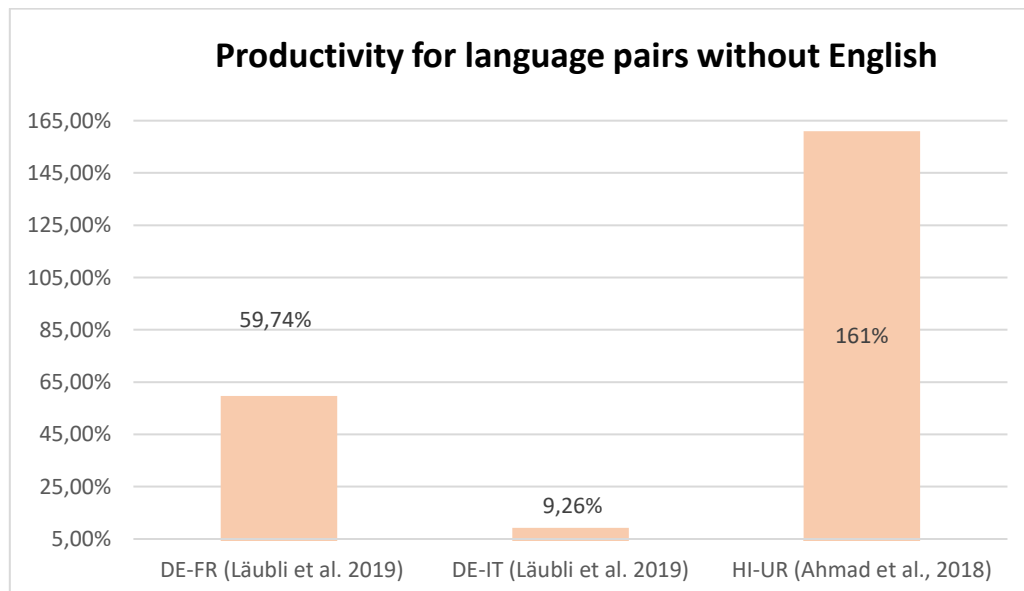


Figure 20: Productivity for language pairs without English

Only three articles examine language pairs that do not include English. The results for them are various: from 9,26% for DE-FR, through 59,74% for DE-IT, to 161% for HI-UR. The first two numbers come from the same study by Läubli et al. (2019), but differ by 50,48% which could

be surprising, since DE-FR and DE-IT have a similar percentage of lexical similarity. HI-UR are closely related, which could be the reason for the high productivity increase.

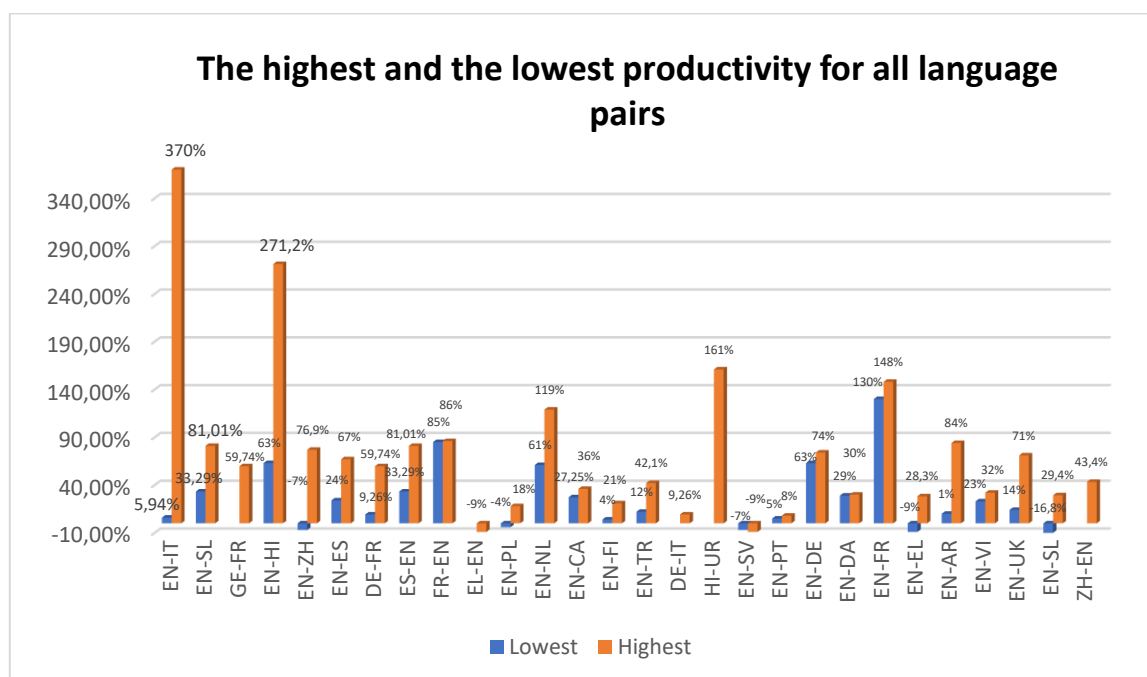


Figure 21: The highest and the lowest productivity for all language pairs

Overall, it is evident that productivity gains vary widely depending on the language combination. Surprisingly, the highest score comes from an EN-IT and EN-HI post-edited text, despite them belonging to different language families. On the third place of the highest scores, is the language pair HI-UR. The lowest scores were reported for EN-SL (-16,8%). In the case of EN-SV, as well as EL-EN no gains were recorded (-7% and -9%, and -9%, respectively). Some language pairs visibly show a tendency to higher productivity gains than others. Some language pairs visibly demonstrate a tendency toward higher productivity gains than others. It is worth mentioning that although translating in one direction may yield significant productivity gains, the same effect is not guaranteed when translating in the opposite direction. Due to the limited number of articles for some of the language pairs, further exploration is needed to confirm the analyzed outcomes.

4.2.1.2 Productivity and types of participants

Figure 22 presents research findings on productivity gains and/or losses for professional translators, while Figure 23 displays numbers for translation students.

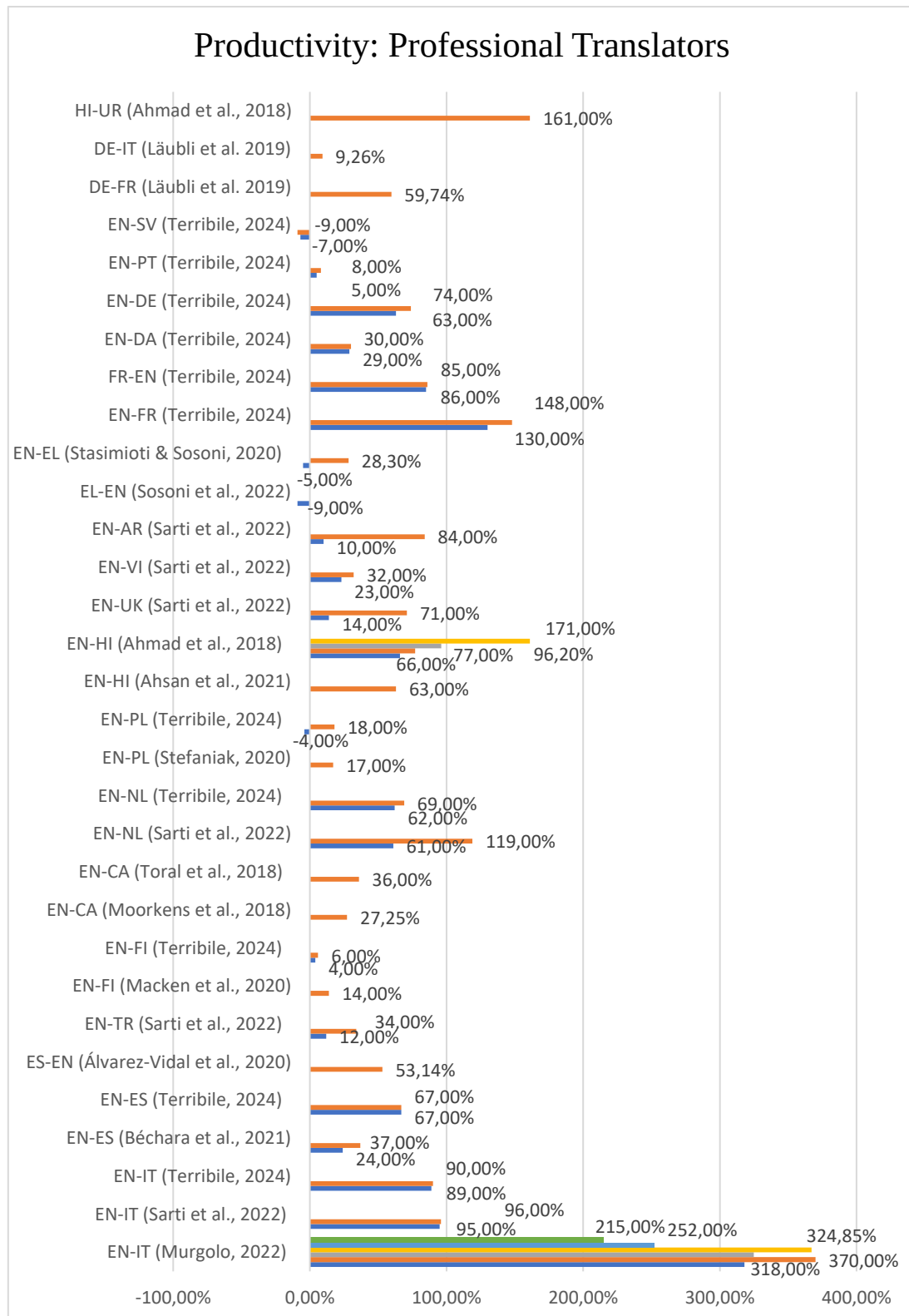


Figure 22: Productivity of professional translators

Various scores within a study in the figure above refer to a scenario with and without outliers (Terribile, 2024), various text domains (Murgolo, 2022) (see Chapter 4.2.1.5 Productivity and text types, different machine providers (Sarti et al., 2022) (see Chapter 4.2.1.4

Productivity linked to NMT providers, MTQE and no MTQE (Bechara et al., 2021), different levels of translation experience (Ahmad et al., 2018).

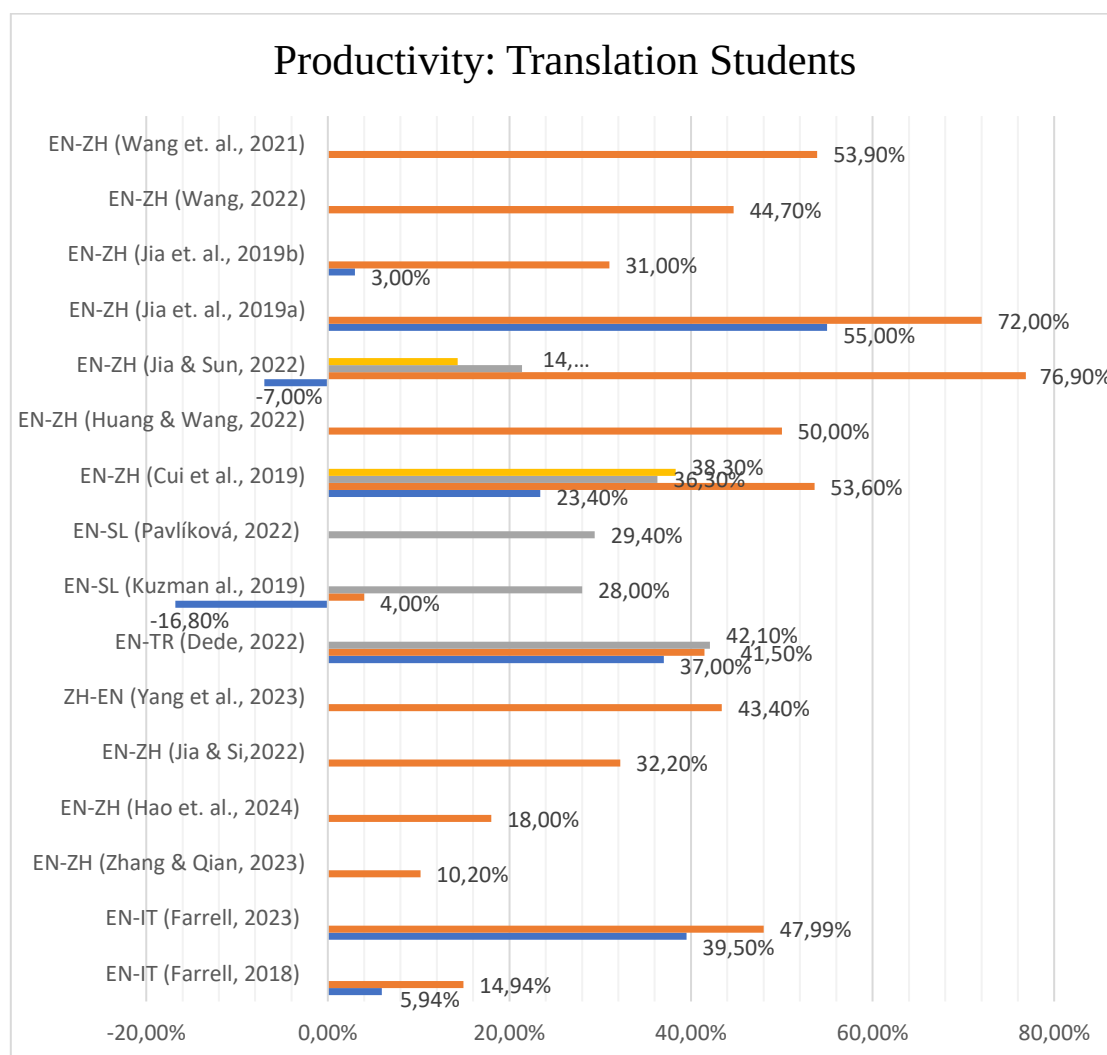


Figure 23: Productivity of translation students

If a study provided more than one value visible in Figure 23, they refer to: different text types/domains (Jia et al., 2019a; Jia et al., 2019b; Jia & Sun, 2022; Cui et al., 2019; Farrell, 2023; Farrell, 2019) (see Chapter 4.2.1.5 Productivity and text types), NMT providers (Kuzman et al., 2019; Dede, 2022) (see Chapter 4.2.1.4 Productivity linked to NMT providers).

As can be seen above, productivity gains and losses for translators vary from -9% to 370%, and from -16,8% to 76,9% for translation students. Overall, the average productivity gains for translators are higher (four studies report gains over 115%), although more losses (coming from five different scenarios) were recorded in this group. Only two losses were recorded for students. Gains and losses for the group presented in Figure 22 vary much more than the ones in Figure 23, that are more consistent.

It should be noted that the line between the two groups mentioned above is sometimes very fine. For example, if a student had just graduated, they were described as professional translators, however, if they were about to finish university, they were assigned to the group of students. Of course, the group of professional translators also included translators with many years of experience, nevertheless, students participating in the studies were usually final-year translation students. Only in the research conducted by Wang (2022), participants were not translation, but English language students, who had had basic transition courses. This group, however, did not show worse productivity than translation students, the opposite – they worked faster than many translation students, however the quality was not measured.

In several instances, the PE productivity did not surpass ten percent, occurring in ten cases in the group of translators, while in the group of students, this was the case six times (that includes two productivity losses for students, and six for professionals). It could be due to a larger amount of research conducted among professionals, which simply brings more results, or due to the fact that students are often more aware of technology and started using MT earlier in the process of preparing for their profession. As a result, while they may not possess the same level of skill as professional translators, they can benefit from PE in terms of productivity, as it handles a substantial part of the task for them. The related danger is the priming effect – it may seem that they work faster, however, the quality of their translation products may decrease. This aspect will be discussed in Chapter 4.2.2.2 **Quality and types of participants**.

In summary, while professional translators typically demonstrate higher productivity levels than translation students, their status does not automatically ensure PE productivity gains. Other factors seem to play a much bigger role, as evidenced by the significant variations in scores, not only across different studies but also within the same study. On the other hand, when provided with suitable circumstances, professional translators perform much better than translation students, which can be seen in the productivity gains – the highest gain for professionals was 370%, whereas for students 76,9%. Then again, students' texts show less losses, however it could be resulting from the uneven number of articles, in which students and professional translators participated.

4.2.1.3 Productivity and post-editing experience

Data available below refers to productivity of participants with PE experience (Figure 24), without PE experience (Figure 25), and to studies in which it was not possible to separate the

data for participants belonging to both groups (Figure 26). Participants who received PE training before the experiment were also put into the second category – that was the case for Jia and Sun (2022), and Sosoni et al. (2022), as well as Jia and Si (2022). The studies conducted by the following researchers did not provide information about the participants’ experience; therefore, they could not be included in this section: Toral et al. (2018) and Wang (2022).

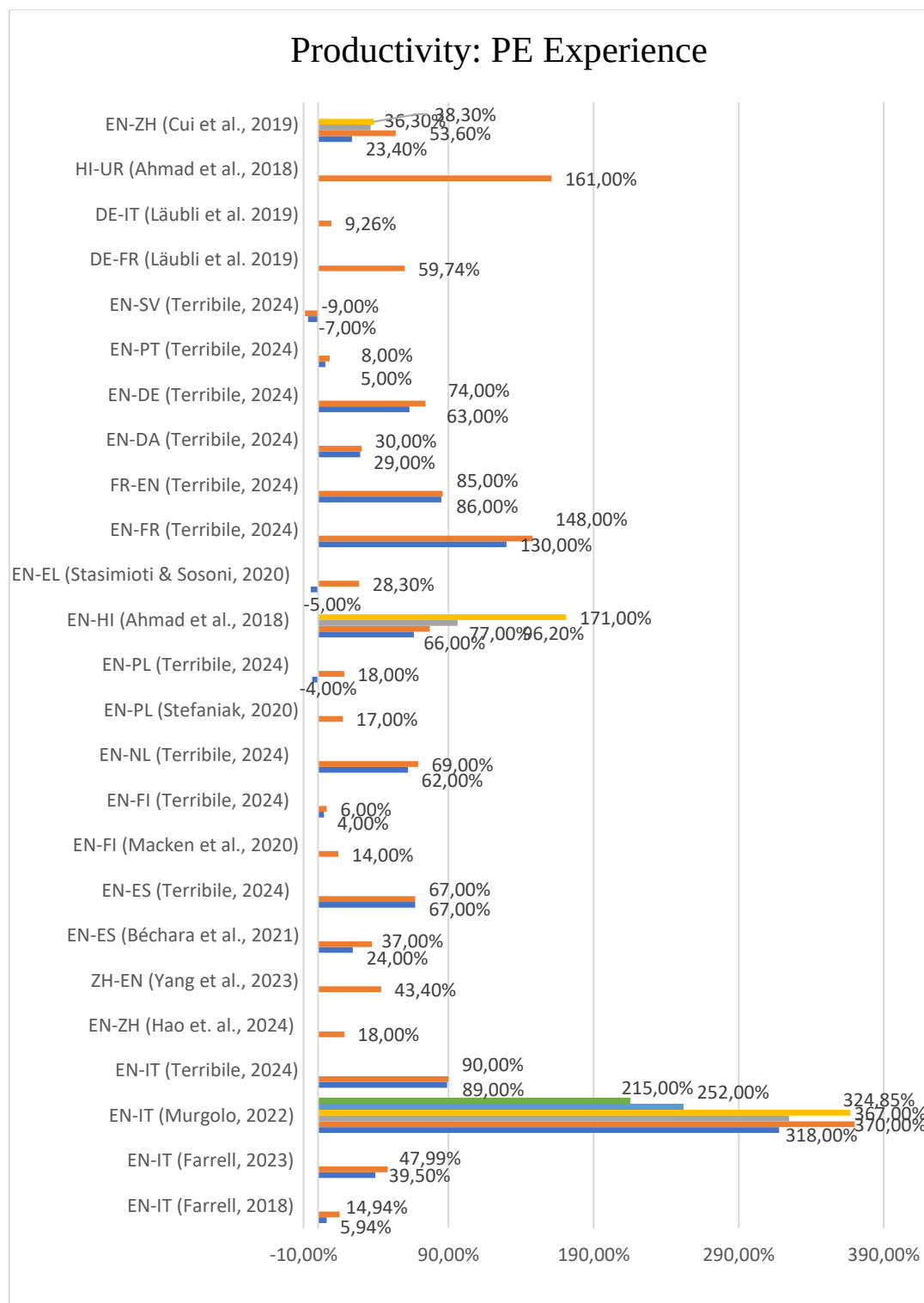


Figure 24: Productivity of participants with PE experience

Various scores within a study in the figure above refer to: a scenario with and without outliers (Terribile, 2024), various text domains (Cui et al., 2019; Farrell, 2023; Farrell, 2019; Murgolo, 2022) (see Chapter 4.2.1.5 Productivity and text types), MTQE and no MTQE (Bechara et al., 2021) different levels of translation experience (Ahmad et al., 2018).

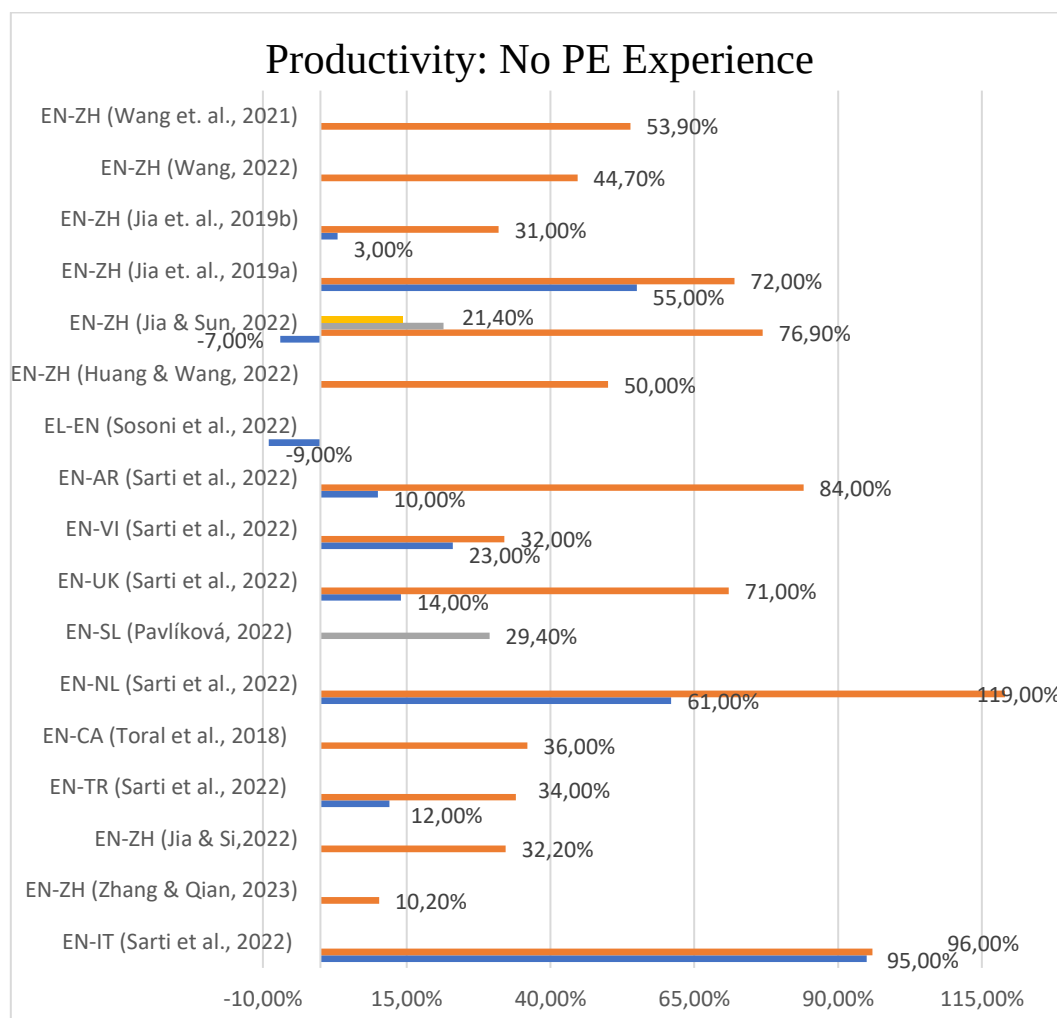


Figure 25: Productivity of participants without PE experience

If a study provided more than one score, they refer to: different machine providers (Sarti et al., 2022) (see Chapter 4.2.1.4 Productivity linked to NMT providers), text domains (Jia et al., 2019b; Jia et al., 2019b; Jia & Sun, 2022) (see Chapter 4.2.1.5 Productivity and text types).

Looking at the figures above reveals that the highest productivity gains were scored by participants with PE experience (370%). This score refers to PE segments in a CAT-tool, as compared to translated segments. It does not mean that the gains from Figure 25 were insignificant, because about half of them reached over thirty percent, and the highest gain equals

119%. Meanwhile, more than half of the research shown in the first figure has reached that number. The lowest values for the first and second category are -7% and -9%, respectively. The first category includes most of the highest scores as well as the lowest. It should not be overlooked that there are quite significant differences, when looking at scores within individual studies ((e.g. Sarti et al. (2022), Jia et al. (2019b), Jia and Sun (2022)). As mentioned before, participants who received a short PE training right before experiment, also show inconsistent scores – in research of Sosoni et al. (2022), it was a loss of 9%, Jia and Sun (2022) describe productivity decreases and increases between -7% and 76,9%, whereas Jia and Si (2022) a gain of 32,2%. This shows that participants with similar experience can perform very differently. Interestingly, the study by Murgolo (2022), described in Figure 24, provides information about the level of experience and emphasizes that the group with the lowest increase in productivity—215% and 252%—had the least experience. While these gains are still impressive, they fall over 110% short compared to the highest gains recorded in the same study.

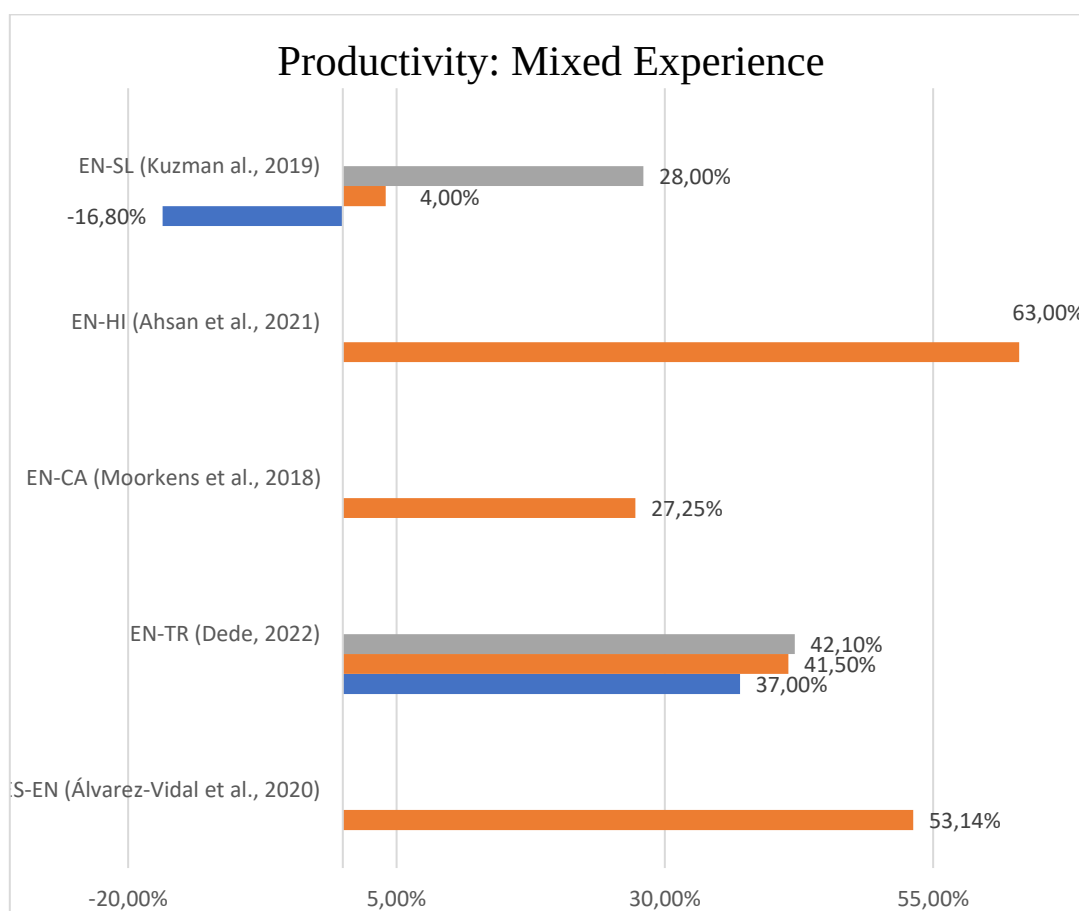


Figure 26: Productivity of participants with mixed experience

Different scores in research conducted by Kuzman et al. (2019) and Dede (2022) refer to various qualities of NMT systems (see Chapter 4.2.1.4 Productivity linked to NMT providers).

Figure 26 shows results for studies in which participants had both PE experience and no experience. In the study conducted by Ahsan et al. (2021), 70% was experienced in PE and 30% not. Alvarez-Vidal et al. (Álvarez-Vidal, et al., 2020) report that 42% has gathered some experience, for Moorkens et al. (2018) it was 33% (2 out of 6) and in case of Dede (2022) 25% claimed to have posted-edited before. 83% of translators participating in the study conducted by Kuzman et al. (2019) had PE experience.

The research that shows the lowest numbers is the only one in which the participants were students, even though they had the highest percentage of PE experience. The study tested very different NMT systems with the same participants, so the lower numbers could also be related to the quality of the NMT systems. Nevertheless, the highest gain for the best performing system, still was not the best visible in Figure 26. Dede (2022), for instance, shows scores for different qualities of the same system (GT); therefore, the differences are minor.

The other two studies, in which the percentage of experience was the highest, and in which professionals took part, have shown the highest productivity gains in this category. Specifically, Ahsan et al. (2021) report a 63% increase, while Álvarez-Vidal et al. (2020) report a 53.14% increase.

Studies presented in Figure 26 do not show individual gains and losses of the experienced and unexperienced participants. Including such figures would be very useful because, first of all, it would provide comparable differences between these two groups of translators, and second of all, contribute to exploring the role of PE experience. An example of such a study is the research presented by Stasimoti and Sosoni (2020), which provided both average and individual scores for translators, with and without experience.

All in all, productivity increased significantly more for participants with PE experience, although the second group also achieved impressive results, thus being skilled in PE does not necessarily mean better results. The lowest average productivity gains were reported in studies where participants had mixed PE experience. At the same time, these studies reported no loss in productivity. Scores vary not only within both groups, but also within specific studies. It is proved by productivity decreases recorded in both categories. This conclusion finds its confirmation also in the studies in which participants, with the exact same experience, achieved different productivity scores.

4.2.1.4 Productivity linked to NMT providers

Figures 27 and 28 consist of data for studies in which PE was performed on Google Translate and DeepL translated texts.

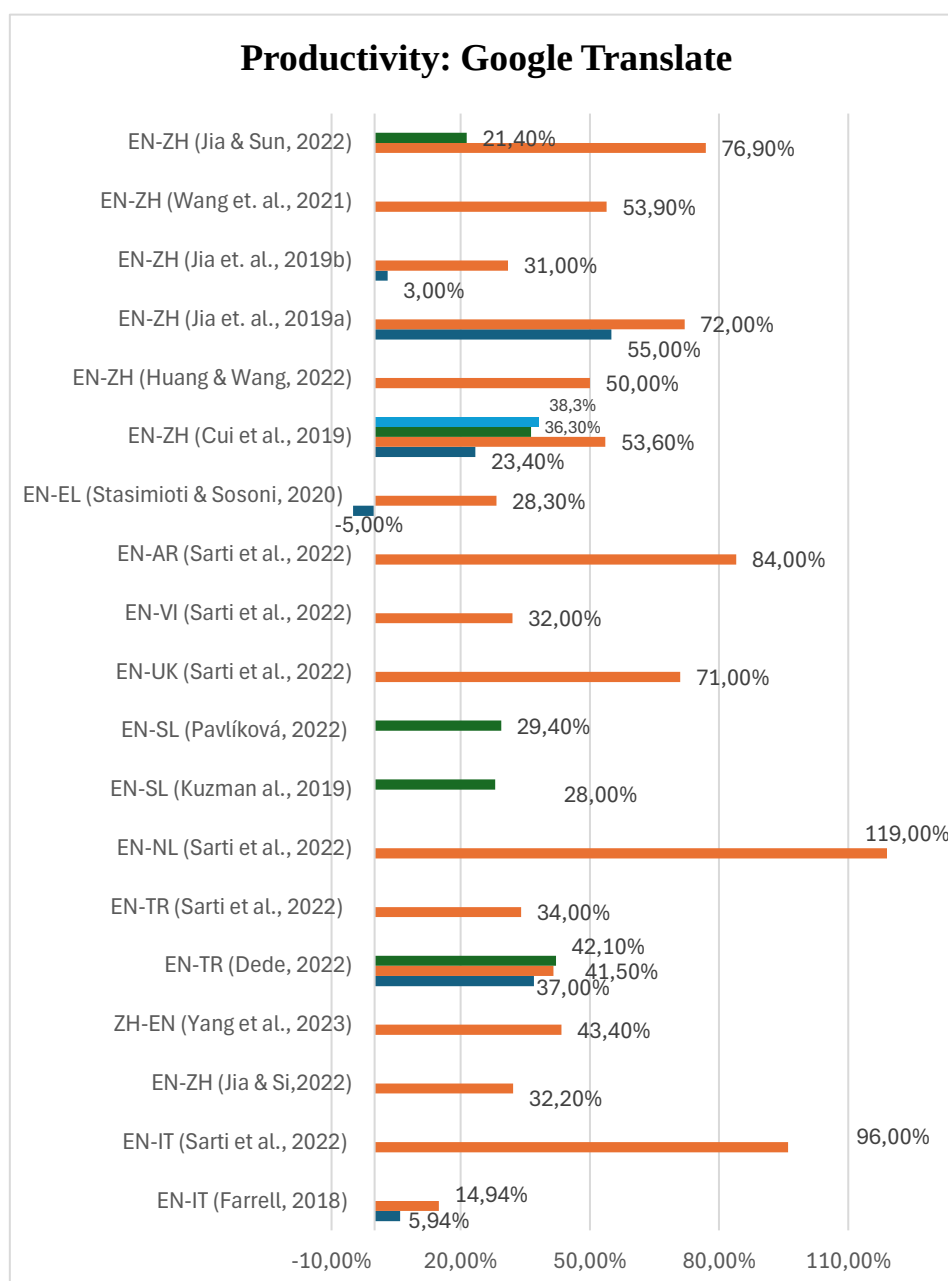


Figure 27: Productivity when PE Google Translate output

Various scores within studies visible in the figure above refer to different texts (Cui et al., 2019; Farrell, 2018; Jia & Sun, 2022; Jia et al., 2019a; Jia et al., 2019b) (see Chapter 4.2.1.5 Productivity and text types), levels of experience (Stasimioti & Sosoni, 2019), qualities of NMT system (Dede, 2022) (see Chapter 4.2.1.4 Productivity linked to NMT providers).

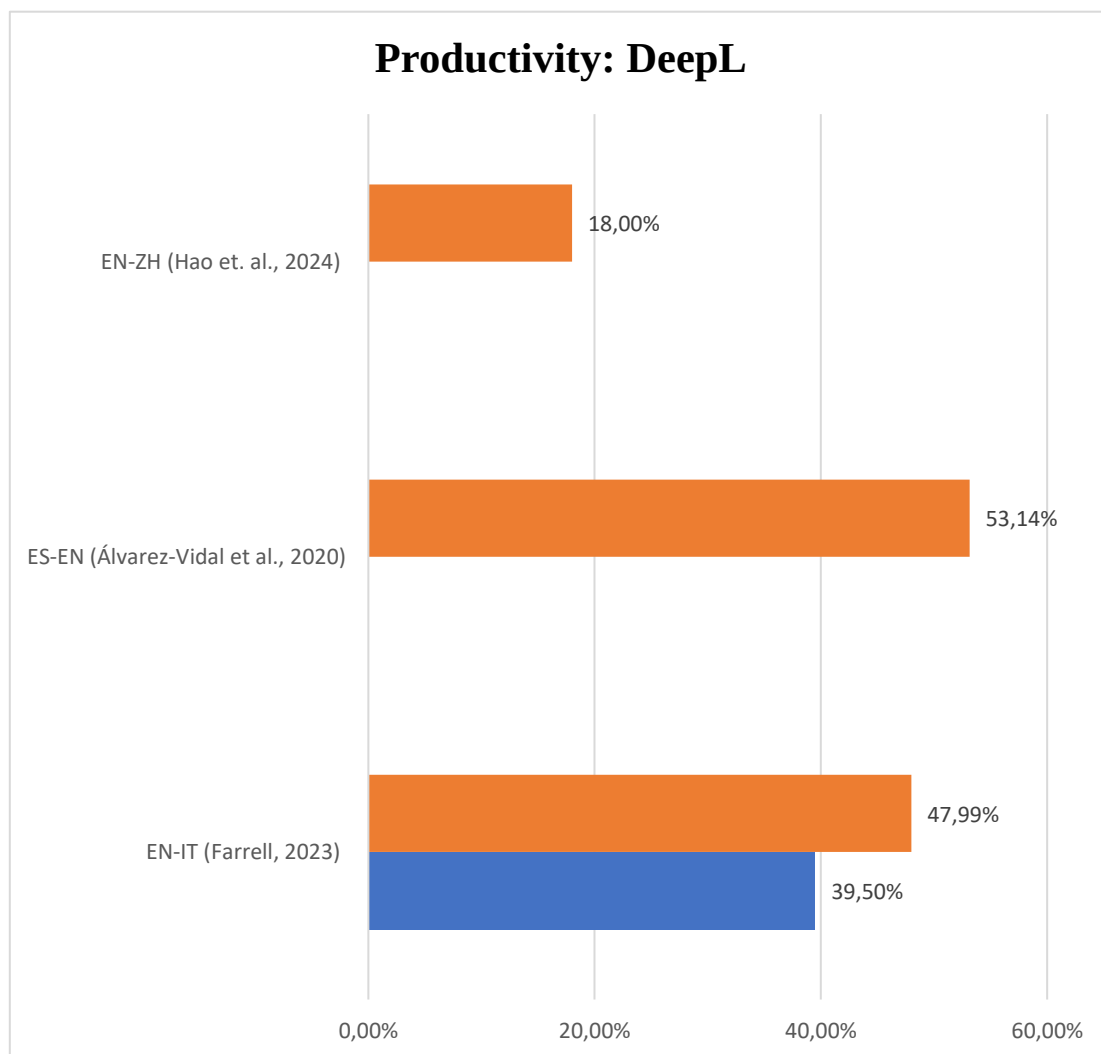


Figure 28: Productivity when PE DeepL output

In the figure above, Farrell (2023) discusses gains for two different texts (see Chapter 4.2.1.5 Productivity and text types).

Google Translate shows mixed results, from -5% productivity decrease to 119% productivity increase. It is worth pointing out that 12 studies for Google Translate, starting from the year 2022, report gains of at least 21%. Moreover, the research starting in the year 2023 reports stable gains between 39 and 48%. It shows that PE with GT has been improving.

Gains for DeepL are relatively consistent – the difference between the highest and lowest number is only 35,14%. DeepL did not bring the highest gains, however it can be stated that it gives the highest level of certainty, since no losses were reported. Nevertheless, it must be noted that the corpus of this thesis includes only three articles available for DeepL, hence the reality regarding gains and losses could be different.

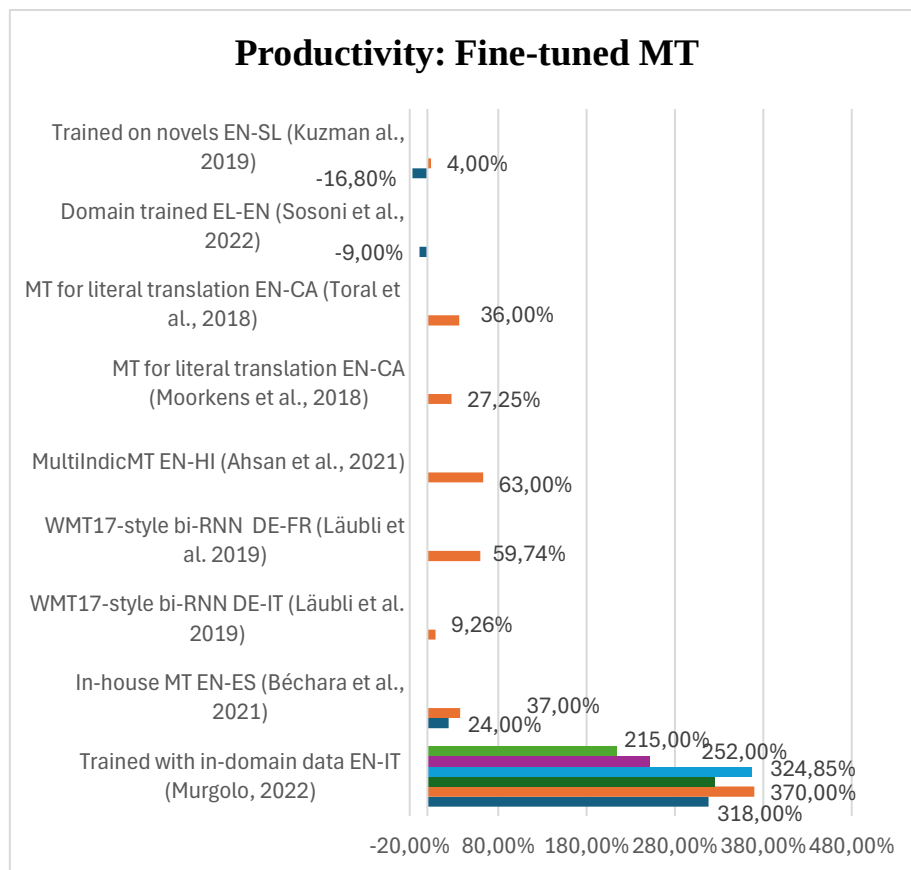


Figure 29: Productivity when PE fine-tuned MT output

Scores showed by Murgolo (2022) describe gains for various text domains (see Chapter 4.2.1.5 Productivity and text types. Kuzman et al. (2019) refer to two NMT systems (see Chapter 4.2.1.4 Productivity linked to NMT providers).

Figure 29 displays data for different NMT engines. As the data above presents, these scores vary strongly from -16,8% to 370%. The lowest score was a result of post-editing for a text translated by a model trained on a novel. This, however, did not bring the desired outcome. In contrast, another system trained on in-domain data has achieved extraordinarily high results, reaching up to 370%. The various scores reported by Murgolo (2022) pertain to different text domains.

Another system designed for literary translation has achieved improvements of nearly 28%. Additionally, a study conducted by Läubli et al. (2019) reported gains ranging from just under 10% to almost 60%. The year in which the systems were trained does not seem to matter – the type of domain or languages should be considered when talking about the productivity of such systems. MultiIndicMT is a system trained on cleaned WAT 2021 for an English-Hindi training corpus. Post-editing a text produced by this system brought the second highest score in

this category. An in-house NMT system trained on the in-house data brought stable gains of 37% and 24% (Béchara, et al., 2021).

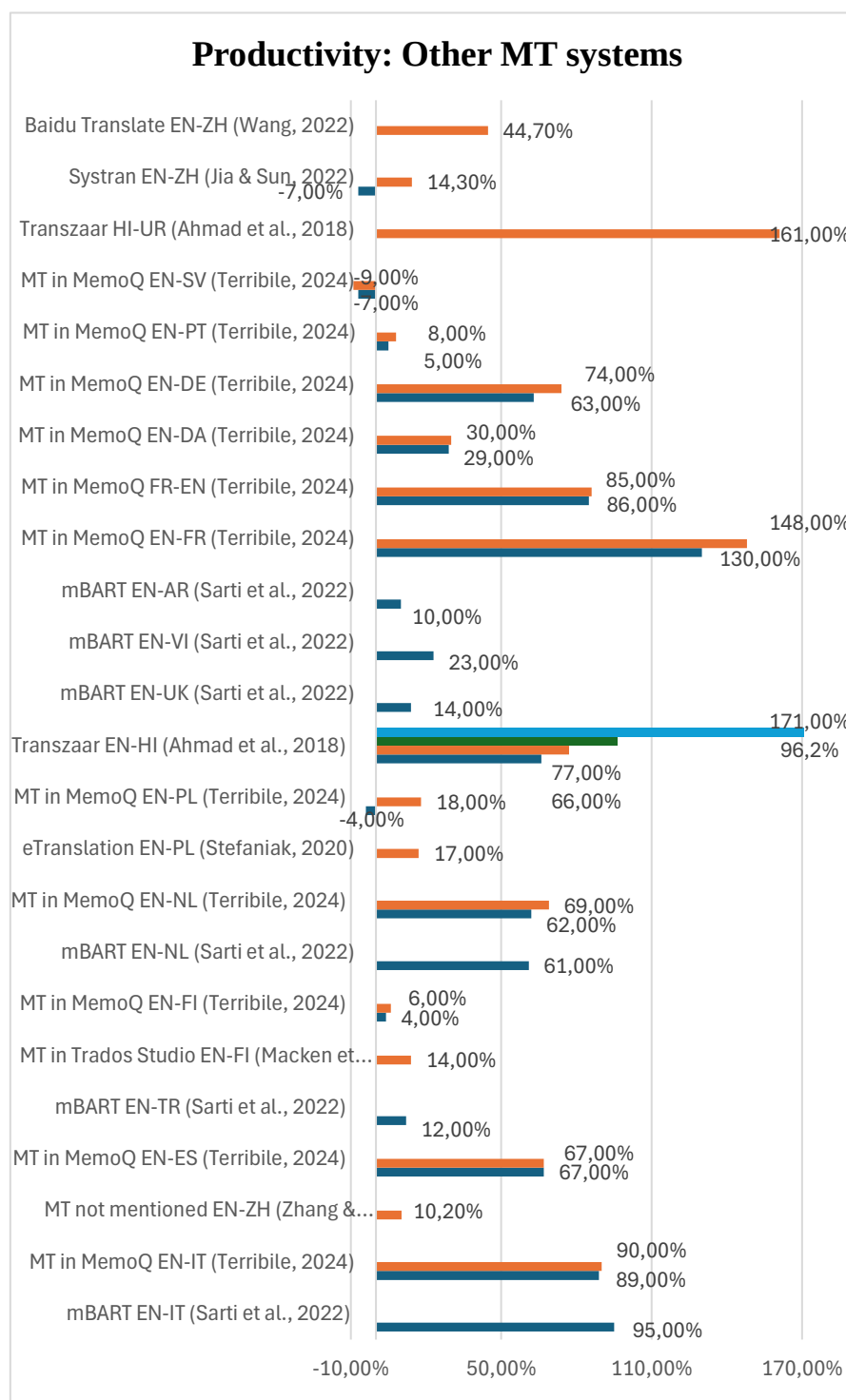


Figure 30: Productivity when PE output of remaining MT systems

In the figure above, various scored presented by authors refer to scenarios with and without outliers (Terribile, 2024), different texts (Jia & Sun, 2022) (see Chapter 4.2.1.5 Productivity and text types), and levels of translation experience (Ahmad et al., 2018).

As can be seen in Figure 30, using Transzaar was the most productive (gain of 171% for expert translators). Transzaar is an AI powered tool that offers computer aided translation (CAT) functionality. In the research (Ahmad, et al., 2018), the participants worked with this tool for a couple of weeks, during which higher and lower gains were recorded, and in which gains were growing, when translators were becoming more comfortable with the software. However, it is important to mention that the gains are compared to results of translating without using any tools. The chart presents the highest reported numbers for this research. NMT system incorporated in MemoQ has brought various results, from gains to losses (Terribile, 2024). In this scenario, the language pair was playing a role. mBART-50 was one of two engines used by Sarti et al. (2022) as a comparison to Google Translate and a representation of state-of-the-art open-source multilingual, even though it usually performs poorer than bilingual systems (Sarti, et al., 2022). Post-editing the output of mBART-50 resulted in productivity gains similar to the ones from NMT systems integrated in CAT-tools (MemoQ (Terribile, 2024) and Trados Studio (Macken, et al., 2020). eTranslation system used by the European Commission brought a 17% productivity gain, although implementing a system used by such an organization sets higher expectations. Baidu Translate, a free NMT system, often used for translating into and from Chinese, produced a text, which post-editing was 44,7% faster than translating from scratch (Wang, 2022). Both tasks were performed in Translog-II. It is a similar result to the one achieved when using Google Translate for the EN-ZH language pair. Systran was chosen as the worst-performing NMT system for EN-ZH, to investigate the differences between it and the best-performing system, which was Google Translate (Jia & Sun, 2022). The experiment resulted in one productivity loss of 7% for Systran and a gain of 14%. It is noteworthy that the loss was reported for a high complexity text and the gain for a low complexity text.

To sum up, various NMT systems were reported in the studies chosen for the corpus. Using commercially accessible NMT systems brings quite consistent results, even if not the highest. Fine-tuning an MT system is not always beneficial, as it sometimes may result in losses, although, if trained successfully, the gains are more than significant. MT systems integrated in CAT-tools yield moderate profits. Generally speaking, decreases in productivity are spread relatively evenly among discussed NMT systems.

4.2.1.5 Productivity and text types

This section discusses the relationship between productivity and various text types. Research texts were organized into the following categories: general, literary, specialized and mixed text types.

General texts include:

- news text (Ahsan, et al., 2021; Cui, et al., 2023; Jia & Sun, 2022; Jia, et al., 2019; Yang, et al., 2023)
- general texts from Wikipedia (Farrell, 2023)
- texts from Wikipedia about Venice and Verona (Farrell, 2018)
- not specified general texts (Álvarez-Vidal, et al., 2020; Wang, et al., 2021)
- general language texts that were promotional brochures for Coca-Cola and Starbucks (Jia, et al., 2019b)
- general sentences (Dede, 2022)
- sentences from Wikipedia with mixed topics (Sarti, et al., 2022)

Various scores within studies visible in Figure 31 refer to different NMT systems (Dede, 2022; Sarti, et al., 2022). Jia et al. (2019a) show gains for texts with different difficulty levels, and Farrell (2023) and Farrell (2018) for various topics. The study by Jia and Sun (2022) refers to both previously mentioned factors. The study by Béchara et al. (Béchara, et al., 2021), provides scores for PE with MTQE and without MTQE.

To the group of literary texts belong:

- a poetic text (Cui, et al., 2023)
- texts by Agatha Christie (Hao, et al., 2024)
- texts with metaphorical expressions (Jia & Si, 2022)
- novels (Kuzman, et al., 2019; Moorkens, et al., 2015; Toral, et al., 2018). Kuzman, et al. (2019) used three different NMT models.

Specialized texts build the most diverse group. There are:

- legal texts (Cui, et al., 2023; Sosoni, et al., 2022)
- an advertising text (Cui, et al., 2023)

- specialized texts from the Accreditation Test for Translators and Interpreters in China (CATTI) (Zhang & Qian, 2023)
- specialized marketing texts (Terribile, 2023), for which the author provided score with O and WO
- an American brochure on composting, more specifically a manual on how to build a home vermicompost (Pavlíková, 2022)
- a banking and finance text (Läubli, et al., 2019)
- educational texts from OER Commons (Stasimioti & Sosoni, 2021)
- a patient information leaflet, and a dishwasher manual (Jia, et al., 2019b)
- mechanical, TLC and electronics patents and life sciences (Murgolo, 2022)
- 260 general sentences from the Autodesk Post-editing Parallel Corpus (Béchara, et al., 2021)

The category of mixed texts includes research in which productivity scores were not specified for each text type and were presented collectively. Text types covered by such studies are as follows:

- legal and general texts (Stefaniak, 2020),
- Hindi news media, English exam books, semi-legal government documents, e-learning websites and documentary subtitles, that were post-edited and translated by participants with various translation experience (Ahmad, et al., 2018),
- data in TPR-DB (Wang, 2022),
- subtitles, legal, drama and documentary texts (Huang & Wang, 2022).

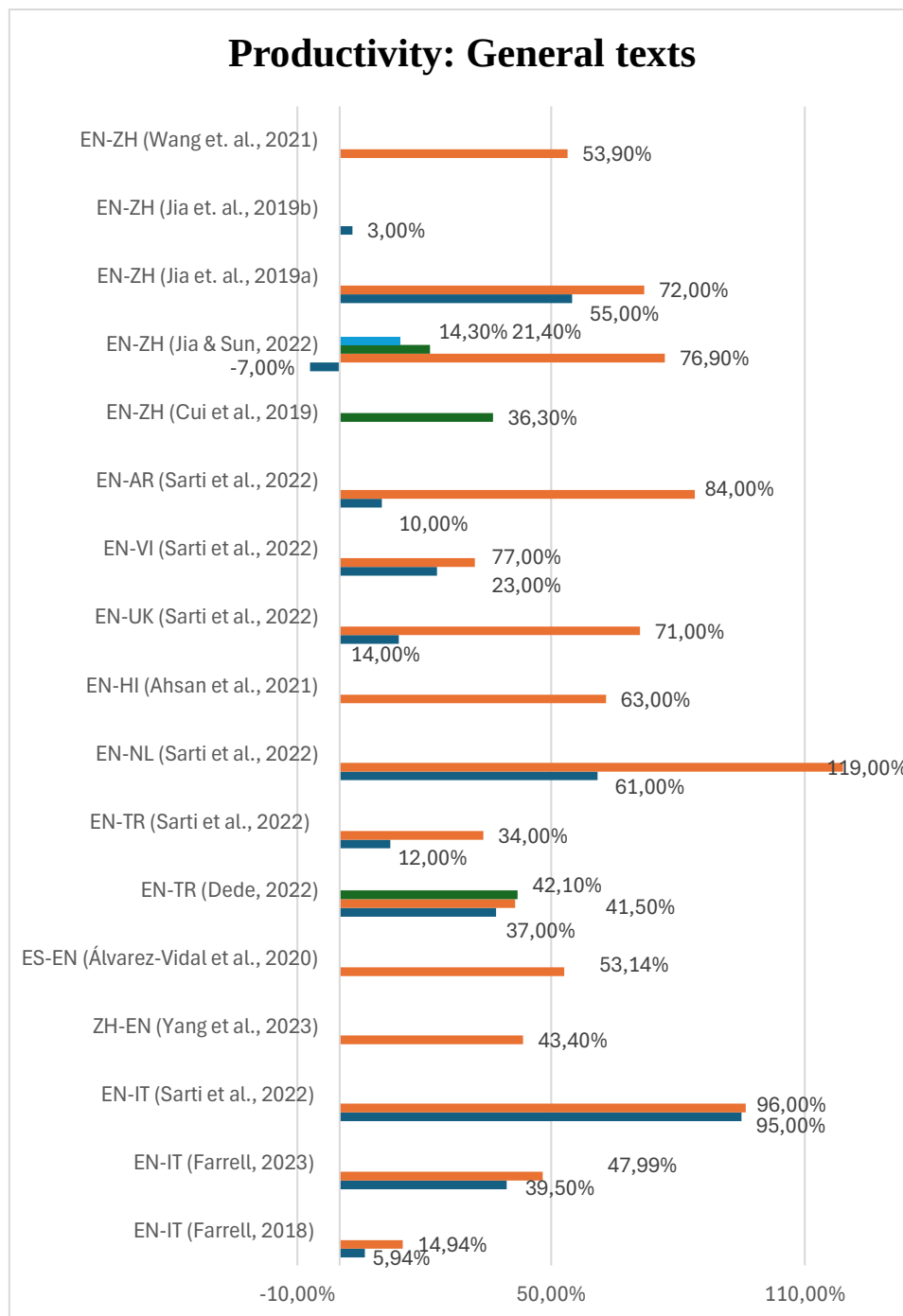


Figure 31: Productivity when PE general texts

Various scores in Figure 31 refer to different MT providers/systems (Sarti et al., 2022; Dede, 2022) (see Chapter 4.2.1.4 Productivity linked to NMT providers), various texts (Jia et al., 2019a; Farrell, 2023; Farrell, 2018) and levels of translation experience (Ahmad et al., 2018).

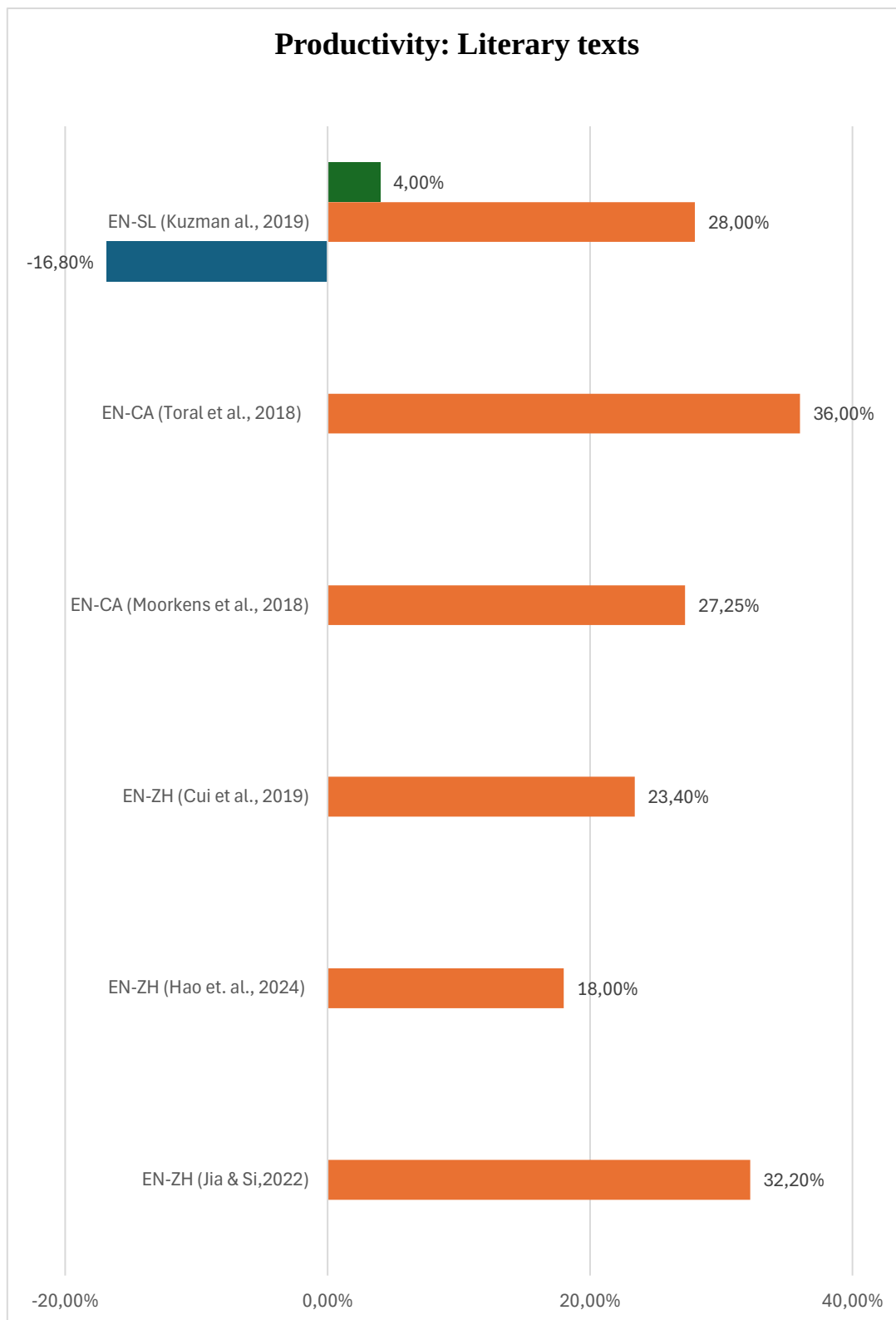


Figure 32: Productivity when PE literary texts

Productivity: Specialized texts

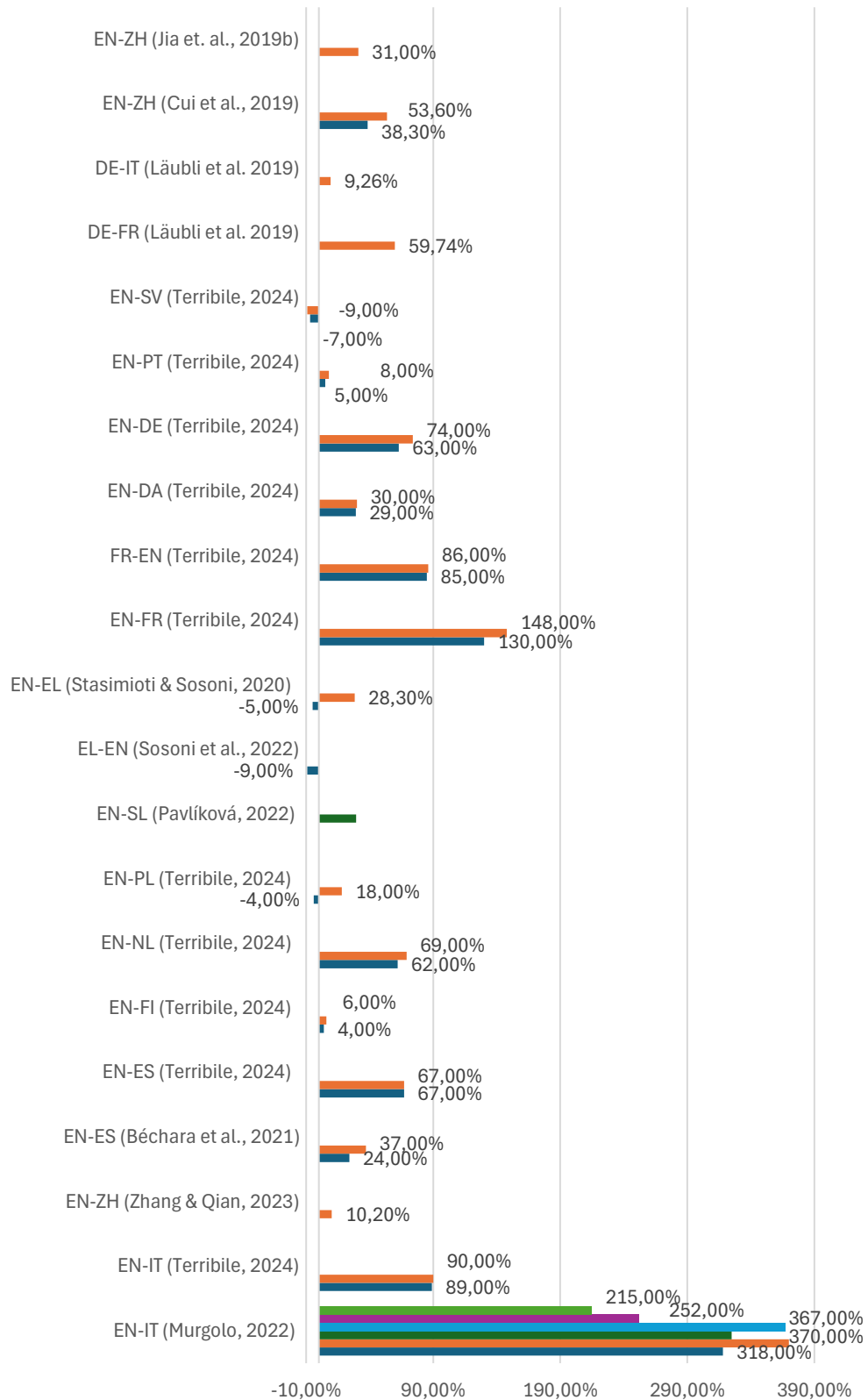


Figure 33: Productivity when PE specialized texts

Various scores in Figure 33 refer to results with and without outliers (Terribile, 2024), various text domains (Murgolo, 2022; Cui et al., 2019), MTQE and no MtQE (Bechara et al., 2021).

The highest percentage values for each category are as follows: 119% for general texts (sentences from Wikipedia on mixed topics), 36% for literary texts (a novel), 370% for specialized texts (mechanical patents). The lowest values for each category are: -7% for general texts (news texts), -16,8% for literary texts (a novel translated with a fine-tuned system), -9% for specialized texts (legal texts). The overall lowest score was reported for a literary text, for a novel-based trained MT system (the highest score from this study was reported for GT, and the second one for a MT system trained on SPOOK), and the highest for a specialized text. It is also worth noting that post-editing a very technical text has resulted in almost four times productivity gains, however, post-editing a legal text has caused a loss. Both types of texts carry serious legal consequences and risks – mistranslating a mechanical patent could potentially result in physical harm, whereas mistranslating a legal text could lead to an action that directly or indirectly affects legal rights.

The only loss recorded for post-editing literary texts was in the case of a novel-trained NMT. The same study reported gains for other NMT systems. Five losses were noted for specialized texts (marketing texts (Terribile, 2024), educational texts from OER Commons (Stasimioti & Sosoni, 2021), a legal text (Sosoni, et al., 2022)), and one for a general text (news) (Jia & Sun, 2022). Overall, general texts scored the highest productivity gains.

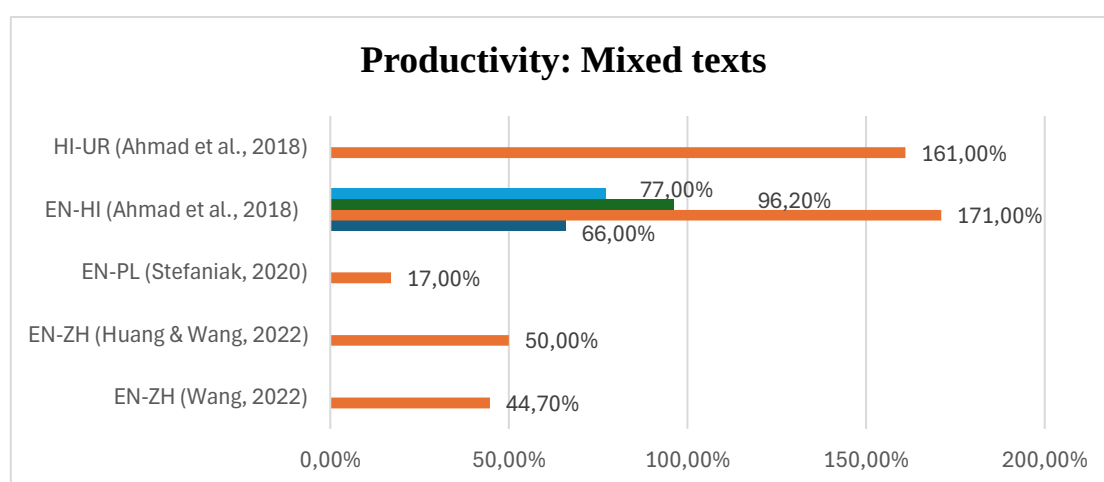


Figure 34: Productivity when PE mixed types of texts

Figure 34 displays the percentages of productivity gains for research that does not specify the scores for each text type that was translated. The biggest gain (171%) was reported in the

research in which the highest number of various text types was translated and post-edited (Ahmad, et al., 2018). The various scores from this research refer to different levels of translation expertise, with the highest being achieved by the most advanced translator. The lowest increase (5%) comes from the second study in terms of textual diversity.

Overall, post-editing specialized texts can be advantageous but also carries certain risks. Notably, post-editing literary texts yielded the most reliable benefits, while general texts achieved the highest overall scores. Therefore, post-editing both general and literary texts appears to be the safest approach.

4.2.1.6 Productivity and experiment settings

Figures 35, 36, 37, 38 and 39 display productivity gains and losses for five types of experiment setting:

- Working in a natural translation environment, which includes a CAT-tool and having access to all types of resources.
- Working in Translog-II (program to record and study human reading and writing processes on a computer), while having the option to utilize a wide range of resources.
- Translating and post-editing in PET (computer-assisted translation tool that supports post-editing).
- Performing tasks in a Word Document, while using other resources.
- Experiments, where translation was done in a Word document, and no resources or very limited resources were permitted.

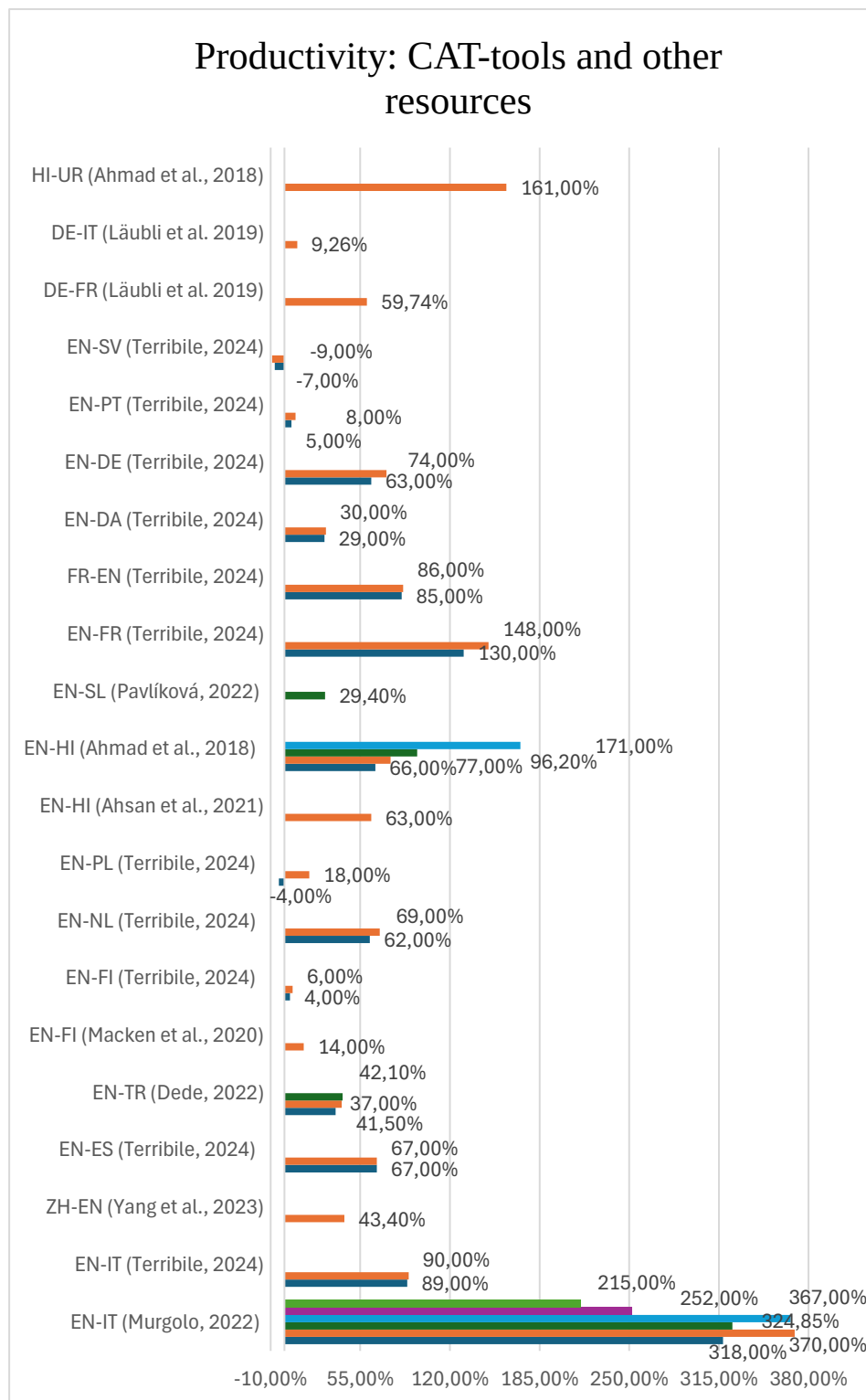


Figure 35: Productivity when PE in CAT-tools and using external resources

In the figure above, scores results reported by Terribile (2024) include or exclude outliers, Ahmad et al. (2018) provide gains for professionals with different translation experience, and Murgolo (2022) explores various text domains, that are described in Chapter 4.2.1.5 Productivity and text types.

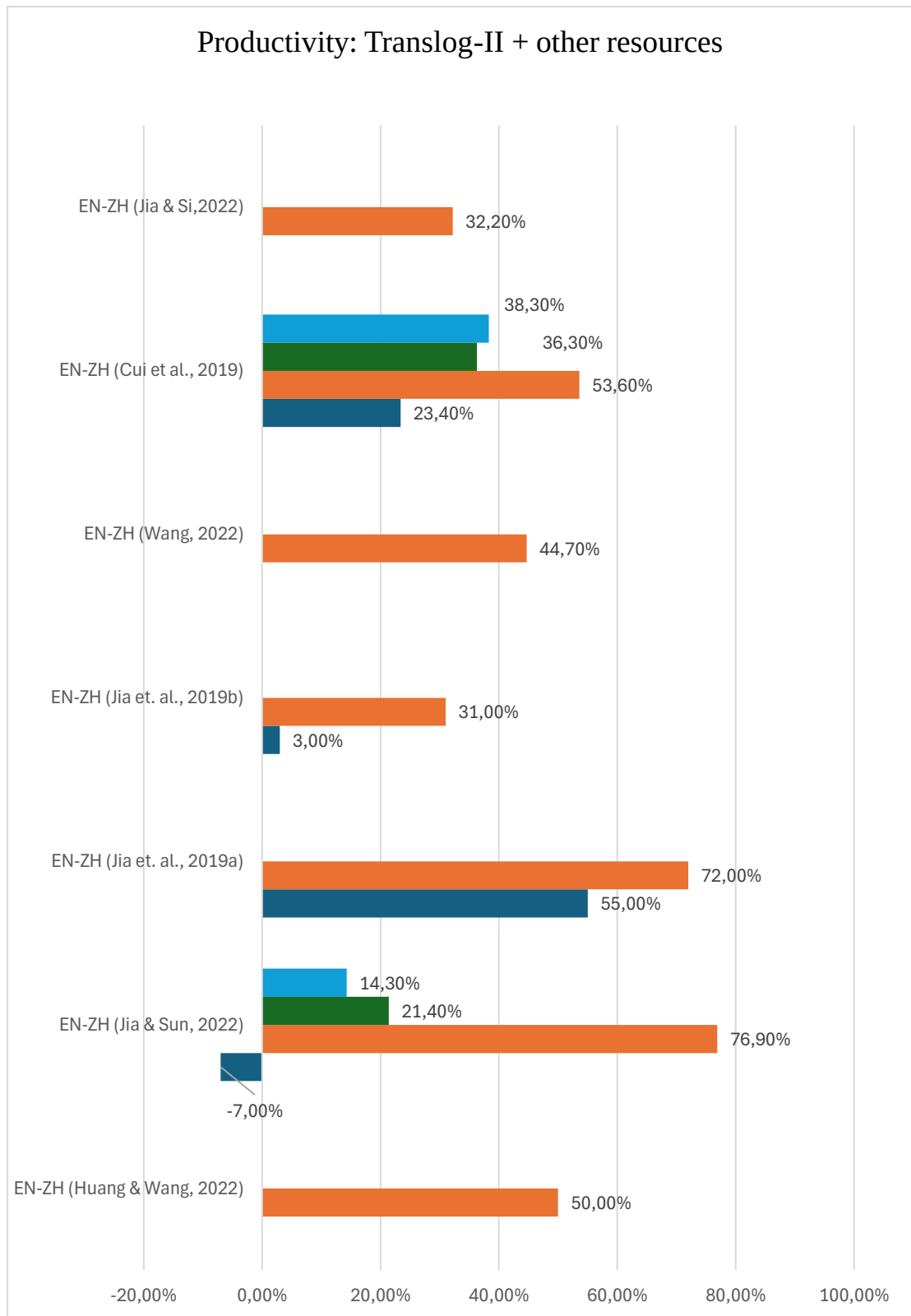


Figure 36: Productivity when PE in Translog-II and using external resources

Studies that provide more than one score in Figure 36 refer to various texts that were described in Chapter 4.2.1.5 Productivity and text types).

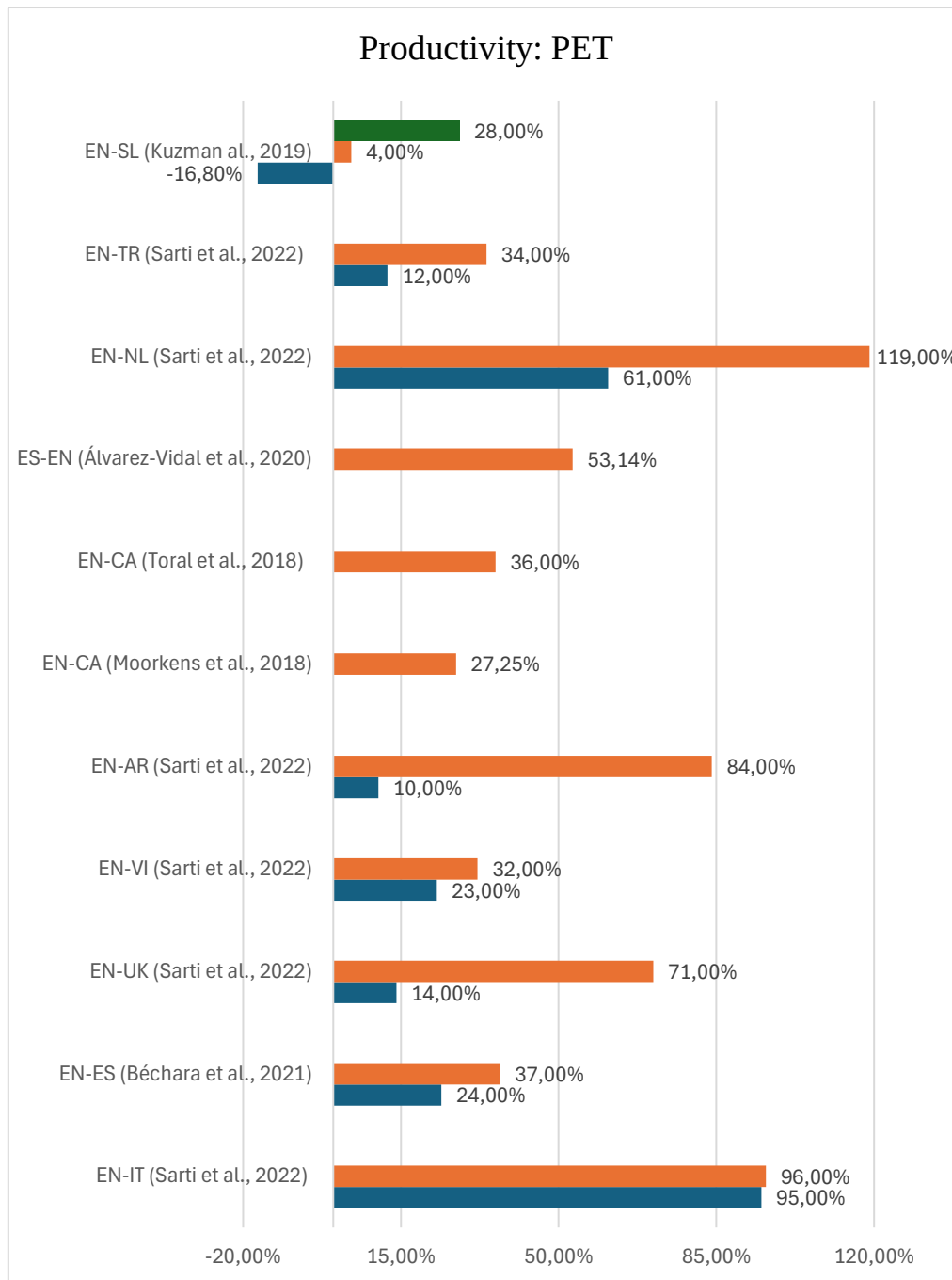


Figure 37: Productivity when PE in PET

Kuzman et al. (2019) and Sart et al. (2022) provide scores for various NMT systems. Details regarding this aspect were described in Chapter 4.2.1.4 Productivity linked to NMT providers).

The figures above reveal that high productivity gains of over 100% appeared in the first three categories. When working in a CAT-tool or/and using other external resources, the highest productivity gain was 370%. It was reported by Murgolo (2022), for participants using a CAT-tool with data base. The lowest loss was 7%, as described by Terribile (2024), for translators working in MemoQ with access to a Translation Memory and MT. The highest and the lowest productivity scores, for translations done in Translog-II, were 76,9% and -7%, both reported by

Jia and Sun (2022). In their research, Eyelink 1000 plus was applied as well. Working in a PET environment resulted in the highest and lowest productivity increase of 119% and -16,8%, as reported by Sarti et al. (2022) and Kuzma et al. (2019). It is important to mention, that the loss in PET setting seems to be related to the NMT system used, since the other two numbers coming from the same study are positive. While participants who used CAT-tools and various external resources achieved the highest productivity gains, they also recorded many of the lowest productivity figures. The variations in gains and losses between PET and Translog-II are similar, with PET performing slightly better.

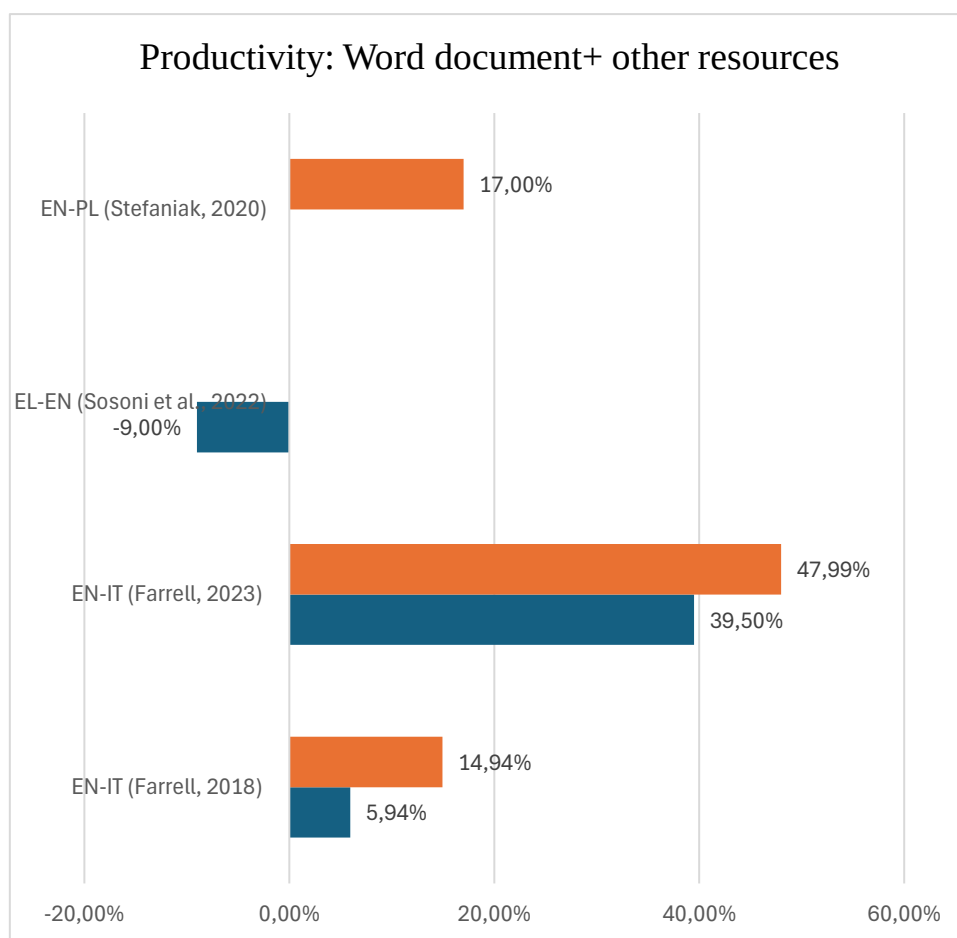


Figure 38: Productivity when PE in a Word Document and using external resources

Studies conducted by Farrell (2018; 2023) show gains for different texts. More details regarding this aspect can be found in Chapter 4.2.1.5 Productivity and text types).

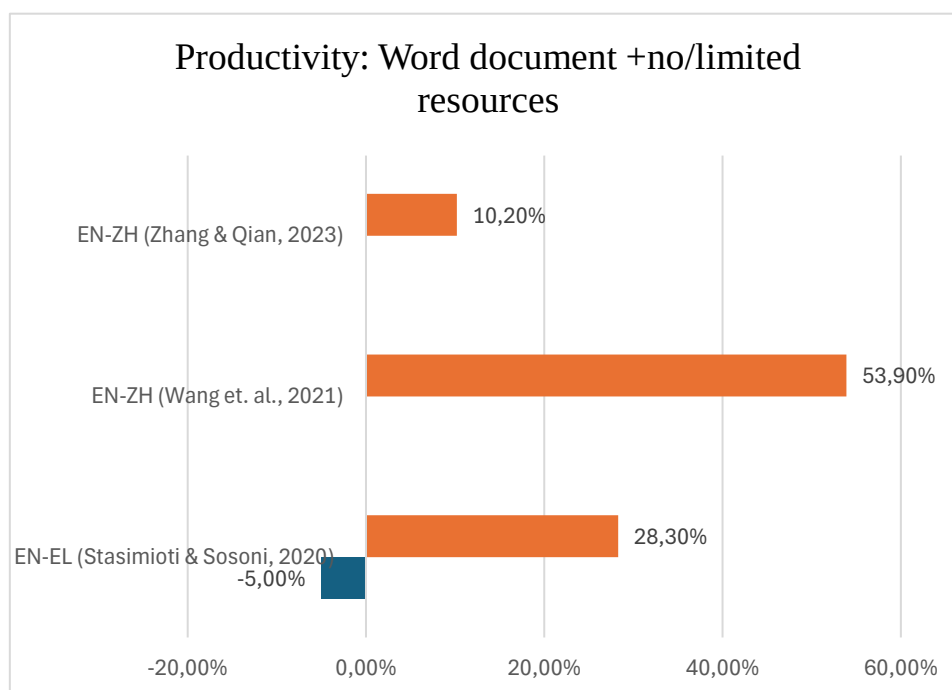


Figure 39: Productivity when PE in Word without any resources or with limited resources

The outcomes of translating and post-editing in a Word document, while being able to access various resources, can range from losses to gains. The same is true for working in Word with limited or no resources.

The biggest loss of -9% occurred in research by Sosoni et al. (2022) and the highest gain in the one by Farrell (47,99%). The three articles belonging to this category applied the exact same scenario.

No online or offline resources were allowed in research of Stasimioti and Sosoni (2021), which resulted in the lowest value in this category. However, the second value reported in this study is much higher. The difference lies in the PE experience; the group with losses lacked experience, while the other group had PE knowledge. Interestingly, the authors focus on the average gain of 19,8%, which could be misleading.

The highest value was reported by Wang et al. (2021). In this scenario, translators were working in Translog-II but had no access to the Internet or any other resources. This comparison shows that even though some translators may not have any resources, they can still be more productive when post-editing. The study by Zhang and Qian (2023) indicates a gain of 10,2% for students who participated in the experiment in a familiar translation laboratory after learning there for a semester. The only resource used was hard-copy dictionaries.

In conclusion, working in a regular translation environment can bring the highest gains. However, it does not always guarantee optimal results, as the lowest losses from the gathered corpus also emerged from this environment. Having limited resources does not always mean

worse results, the opposite, three out of four productivity values were positive. It is important to note that both the lowest and highest values in two of the groups—specifically the PET and Translog-II scenarios—originate from the same study. This suggests that while the working environment plays a role, other factors may have a more significant impact.

4.2.1.7 Chapter summary and conclusion

This chapter has concentrated on the comparative analysis of productivity gains and losses associated with post-edited texts in relation to translations produced from scratch. This examination took into account various factors, including language pairs, participant types, post-editing experience, types of translation systems employed, text types, and experiment settings.

The most significant differences were observed in the EN-IT language pair, while the smallest differences were found in EN-CA. This indicates that languages significantly influence the outcome of translation. Professional translators show more consistent and higher results than students. Experience in PE allows participants to score the highest values, however, it is not unusual for the same group to exhibit varying performance levels. Machine systems yield a variety of outcomes, as gains and losses are distributed across all categories. General texts in PE produced the highest scores overall. Utilizing a CAT tool can lead to both the highest and lowest results. It can be concluded that the majority of the outcomes are individual in nature.

4.2.2 Quality and analyzed factors

This section discusses correlation between quality and factors, such as language pair, types of participants, post-editing experience, types of translation systems, text types, and experiment settings. The following chapter, however, will not be presenting articles in which the quality of PE texts was equal to the HT texts; it will focus on those in which quality gains or losses occurred. Six articles reported quality losses (Ahsan, et al., 2021; Cui, et al., 2023; Farrell, 2018; Hao, et al., 2024; Pavlíková, 2022; Zhang & Qian, 2023), and eight recorded quality gains (Béchara, et al., 2021; Cui, et al., 2023; Jia & Si, 2022; Jia & Sun, 2022; Läubli, et al., 2019; Sosoni, et al., 2022; Yang, et al., 2023; Wang, et al., 2021).

In total, 20 cases of gains and losses were reported, as some research studies provide scores for multiple factors or scenarios. As outlined in Chapter 4.1, certain gains and losses were identified as significant, while others were deemed insignificant. Additionally, some cases

lack sufficient information regarding their significance and are therefore classified simply as gains. For these reasons, quality presented in charts in the next subsections is divided into following categories:

- three insignificant losses (IQL)
- three significant losses (SQL)
- six significant gains (SQG)
- six insignificant gains (IQG)
- two gains (G)

All of the quality scores above will be compared with factors included in this research.

The language pairs included in this section are: EN-ZH, EN-IT, EN-ES, EN-HI, EN-SL. Groups of participants included remained the same — professional translators and translation students, experienced and unexperienced participants. NMT systems described in this section are:

- Google Translate
- DeepL
- in-house MT system
- MultiIndicMT, system trained for EN-HI
- domain-fine-tuned system
- unknown type of MT.

Text types include:

- general texts: news text (Ahsan, et al., 2021; Cui, et al., 2023; Jia & Sun, 2022; Yang, et al., 2023), text about Venice (Farrell, 2018), not specified general texts (Wang, et al., 2021)
- literary texts: poetic text (Cui, et al., 2023), texts by Agatha Christie (Hao, et al., 2024), texts with metaphorical expressions (Jia & Si, 2022)
- specialized texts: legal text (Cui, et al., 2023; Sosoni, et al., 2022), advertising text (Cui, et al., 2023), texts selected from the Accreditation Test for Translators and Interpreters in China (CATTI) (Zhang & Qian, 2023), 260 sentences from the Autodesk Post-editing Parallel Corpus (Béchara, et al., 2021), specialized marketing text (an American brochure on composting; specifically, it was a manual on how to build a home vermicompost) (Pavlíková, 2022), text on banking and finance (Läubli, et al., 2019)

The quality data was available for the following experiment settings: Translog-II + external resources, CAT-tools + external resources, PET and working in Word with limited resources.

4.2.2.1 Quality and language pairs

Figure 40 focuses on comparing quality and eight different language combinations.

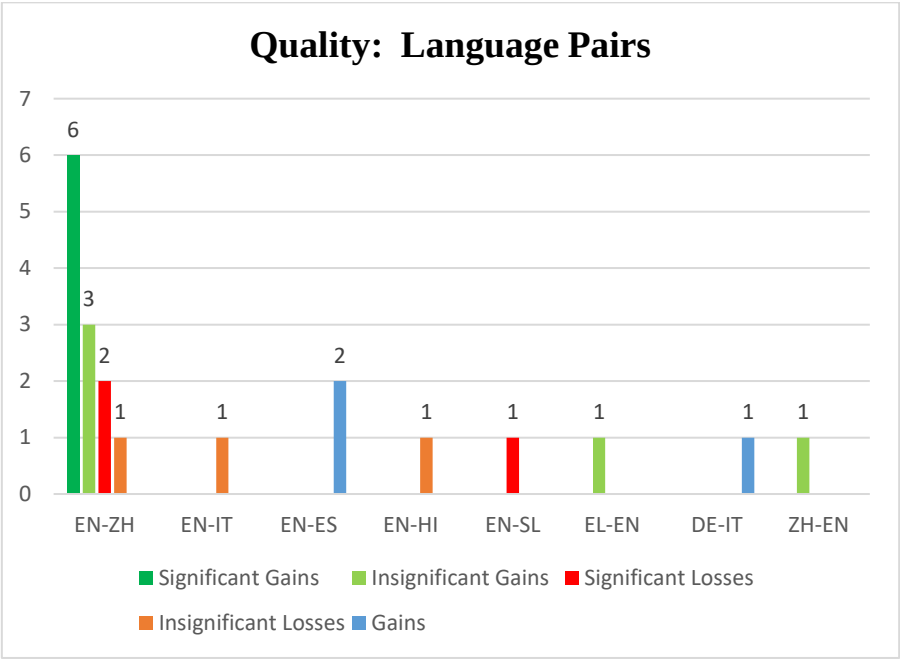


Figure 40: Quality and language pairs

Twelve out of twenty quality scores in the available research were measured in the EN-ZH language pair, for which six out of six SQG, three out of six IQG, two out of three SQL, and one out of three IQL were found. Only one study was available for each of the remaining language combinations included in this section. The only research solely reporting a SQL was conducted for EN-SL. EN-IT and EN-HI show IQL, however, other research in these language combinations assumes no changes in quality (see Ahman et al. (2018), and Farrell (2023) in Table 1).

Summarizing, the most gains (9) as well as losses (3) were recorded for EN-ZH. EN-HI, EN-IT and EN-SL reported losses where for EN-SL the loss was significant. The remaining language combinations show S and I gains. Although some language pairs show visibly more gains/losses, due to an unequal distribution of research, it is difficult to determine the exact influence of language combinations on quality and further research is required.

4.2.2.2 Quality and types of participants

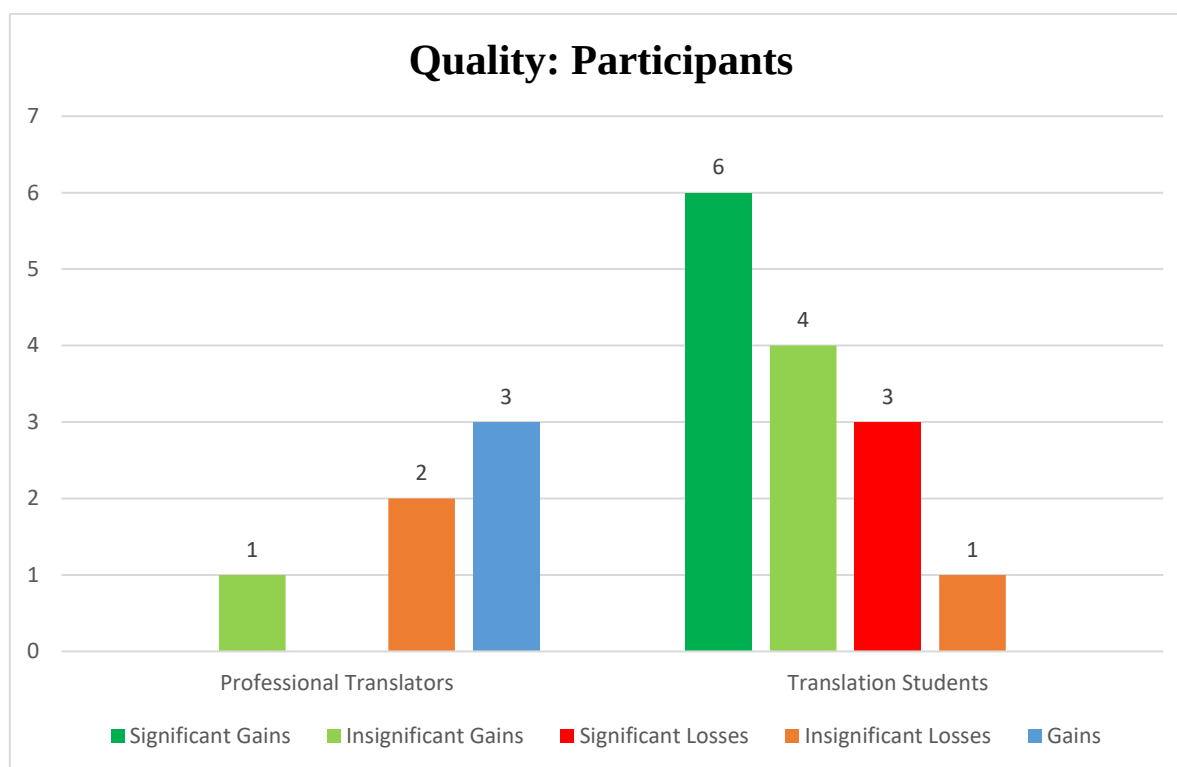


Figure 41: Quality and participants

As visible above, it is possible to distinguish between two groups participating in the analyzed research – professional translators and translation students. Research by Béchara et al. (2021), Ahsan et al. (2021), Sosoni et al. (2022), and Läubli et al. (2019) was conducted among professionals. Students were participants of the following research: Jia & Sun (2022), Farrell (2018), Cui et al. (2019), Wang et al. (2021), Zhang & Qian (2023), Hao et al. (2024), Pavlíková (2022) and Yang et al. (2023). As Figure 23 indicates, the most SQG and IQG were reached by translation students, however, also all SQL occurred in this group.

Regarding quality gains, it seems like translation students have profited from PE more; at the same time, they scored all SQL. It may be a result of the priming effect, as students appear to be more susceptible to machine translation suggestions and are less attentive to potential errors. On the other hand, the lack of SQL on the side of professional translators could be seen as a sign of a better performance. Nevertheless, it should be noted that the data regarding quality changes was available for six cases including professional translators, and fourteen scenarios in which students participated. Therefore, it may be difficult to compare those groups validly.

4.2.2.3 Quality and post-editing experience

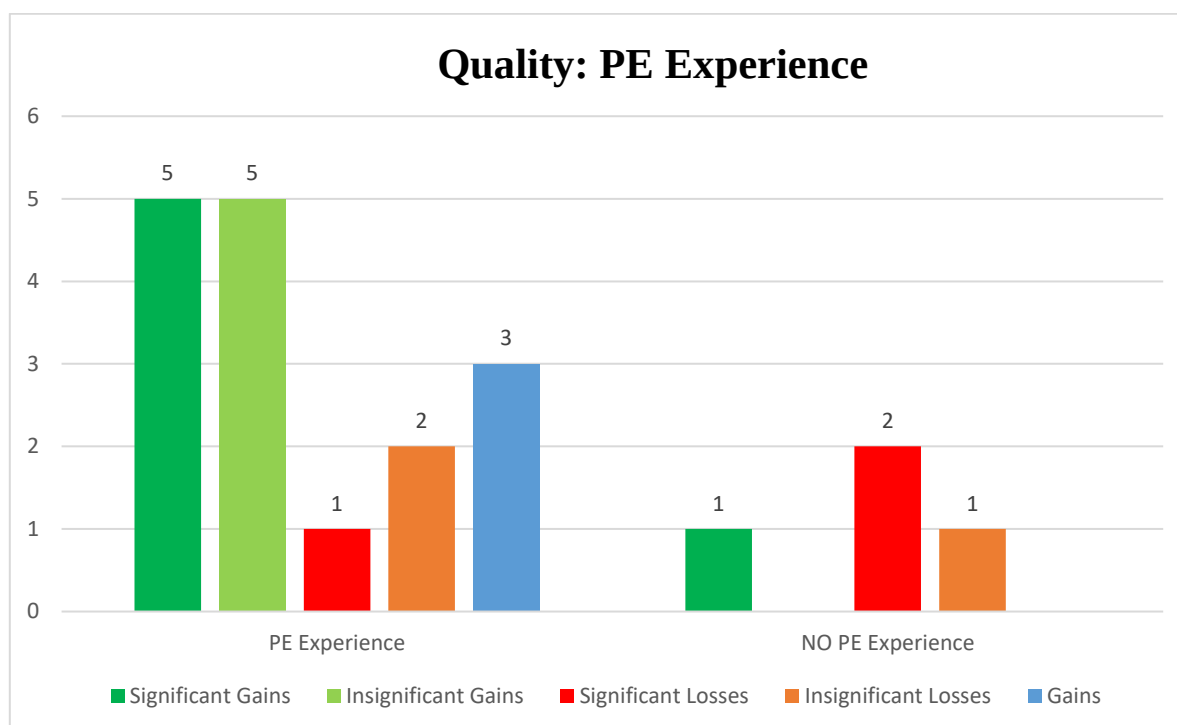


Figure 42: Quality and post-editing experience

The research of Farrell (2018), Cui et al. (2023), Jia and Sun (2022), Sosoni et al. (2022), Hao et al. (2024), Jia and Si (2022), Bechara et al. (2021), Läubli et al. (2019), Yang et al. (2023) invited translators with PE-experience, and the research conducted by Wang et al. (2021), Zhang and Qian (2023), Ahsan et al. (2021) and Pavlikova (2022) created an experiment for translators without PE-experience.

Numbers in Figure 41 illustrate that five out of six SQG and five out of five IQG were achieved by translators with PE experience. Furthermore, among individuals with no prior experience in PE, there were 2 SQL and 1 IQL. For translators who were proficient in PE, there was 1 SQL and 2 IQL.

According to values presented in Figure 41, PE-experience plays a crucial role when correcting MT texts. Therefore, it is a factor that should not be omitted when conducting PE quality and productivity studies. Moreover, the skill of PE in context of MT should be taught at universities and introduced more widely, as it could benefit translators not only in their professional life, but also in the process of preparing for it, by introducing them to a different linguistic task.

4.2.2.4 Quality linked to NMT providers

Six different NMT systems with significant and insignificant gains and losses, as well as general gains can be seen in Figure 43.

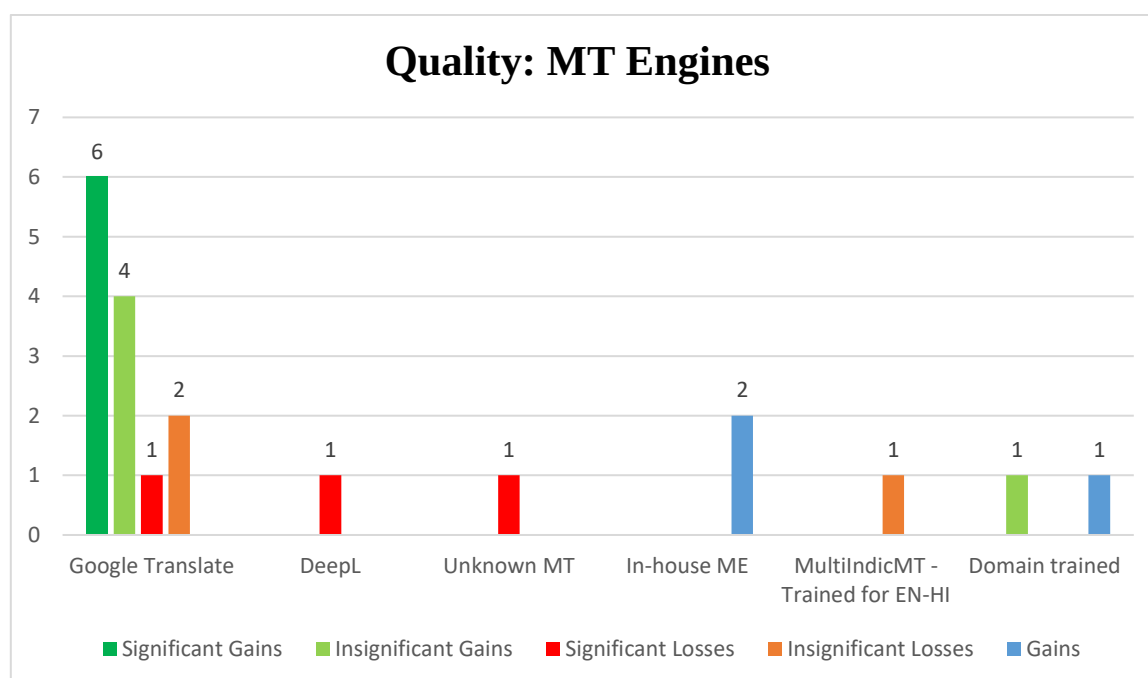


Figure 43: Quality and machine translation systems

Google Translate was the most common MT system used among all thirty articles. Two out of three SQL were produced when using commercial machine systems (Google Translate and DeepL). All SQG came from Google Translate post-edited texts as well. It is impossible to say, which system produced the third SQL, since the authors did not include this information (Zhang & Qian, 2023). Although, surprisingly, domain trained MTE did not bring any SQG, what would usually be expected, no significant losses were caused by them either. Only the MultiIndicMT, a system trained on an English-Hindi training corpus, caused an IQL (Ahsan, et al., 2021). Texts that have been post-edited or translated using an in-house MT system have resulted in either insignificantly or significantly improved quality.

As can be observed from Figure 42, commercial MT systems already bring impressive quality gains. Nevertheless, one must be cautious when using them, since they may also cause serious quality losses. Domain trained and in-house systems bring either slight IQG or IQL. No SQG were reported, which shows that investing time into fine-tuning a machine does not always bring profits. However, having an in-house MT system can produce a text of beneficial quality, that corresponds with the quality guidelines of a company; it seems to be the safest option.

4.2.2.5 Quality and text types

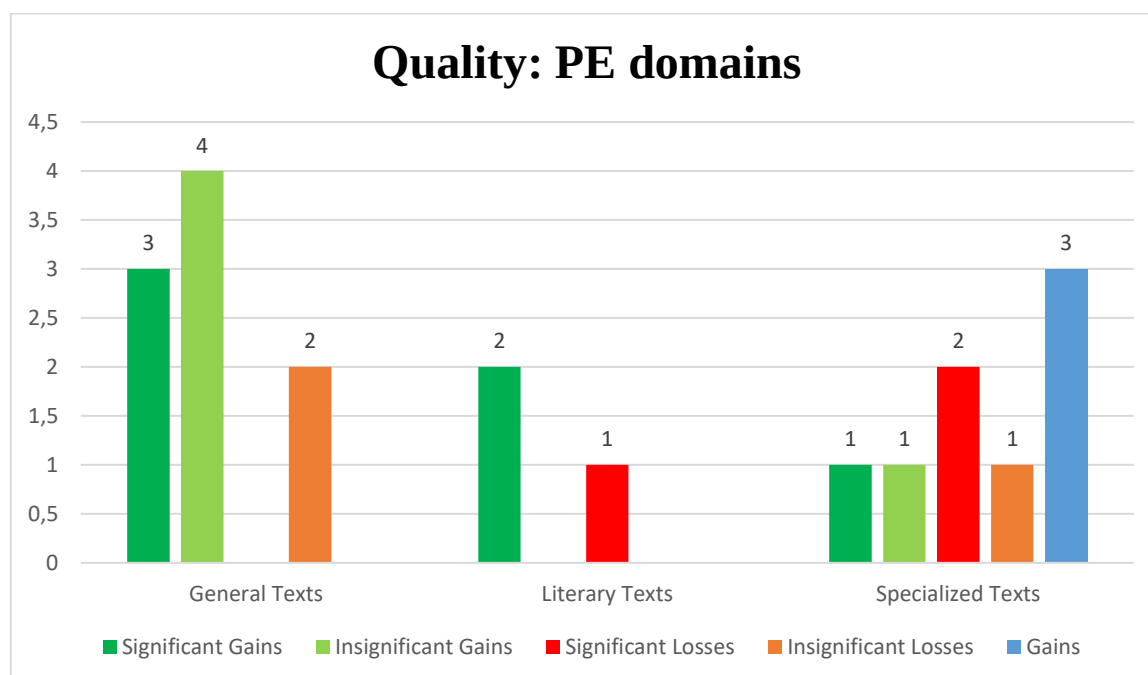


Figure 44: Quality and post-editing different text domains

As can be seen above, PE general texts brought the most SQG and two IQL. SQG were measured for two news texts (Cui, et al., 2023; Jia & Sun, 2022) and a not specified general text (Wang, et al., 2021). IQG in this category were present in three news texts (Jia & Sun, 2022; Yang, et al., 2023). IQL among general texts were reported for a text about Venice (Farrell, 2018) and for news text (Ahsan, et al., 2021).

The remaining two SQG can be found in PE literary texts (texts with metaphorical expressions (Jia & Si, 2022), and a poetic text (Cui, et al., 2023)), along with one SQL (for post-editing Agatha Christie's texts (Hao, et al., 2024)).

The most various results are visible among specialized texts, where numbers from all five quality categories were reported. The domain that scored a SQG was advertising (Cui, et al., 2023). An IQG was spotted in a legal text (Sosoni, et al., 2022). On the other hand, an IQL was reported for a legal text as well (Cui, et al., 2023) and two SQL for a technical marketing text (Pavlíková, 2022) and specialized texts from the Accreditation Test for Translators and Interpreters in China (Zhang & Qian, 2023). General gains were reported for banking and finance (Läubli, et al., 2019) and sentences from the Autodesk Post-editing Parallel Corpus (Béchara, et al., 2021).

To sum up, using PE is the most beneficial for general texts, namely news. No research showed any significant danger to the quality of these texts. Using PE for literary texts can have

a positive as well as detrimental effect, the more literal the text is. Regarding specialized texts, PE brought the best results for an advertising text, however the influence of PE is increasingly negative as the text becomes more technical.

4.2.2.6 Quality and experiment settings

Figure 45 displays the correlation between quality and different experiment settings.

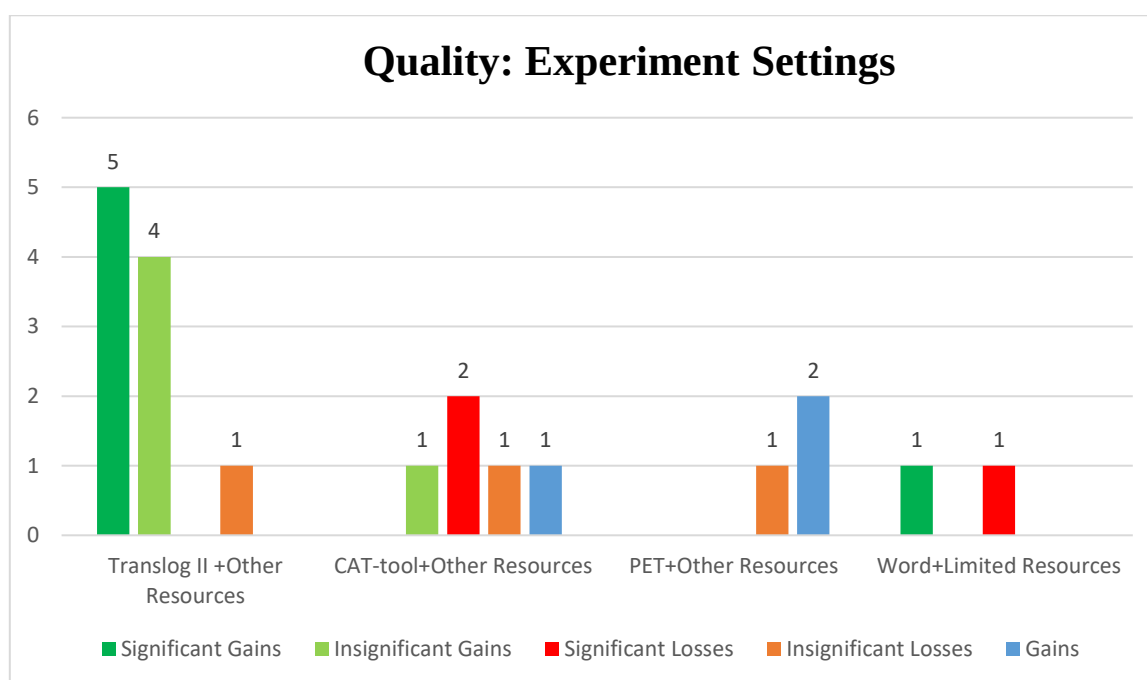


Figure 45: Quality when PE in different experiment settings

As can be observed from the figure above, PE in Translog-II and having access to external resources brings the best outcomes in terms of quality. Five out of six SQG, four out of five IQG and only one IQL were recorded in this category. Mixed results are visible when working in a CAT-tool and using usual resources. Two SQL, one IGL and only two IQG occurred in this scenario. Post-editing in PET produced more gains than losses, whereas working in Word with limited resources caused two opposite outcomes: one SQG and one IQL. The SQG was showed by Wang et al. (2021). In this study, participants were working in Translog-II but had no access to the Internet or any other resources. The SQL comes from a study conducted by Zhang and Qian (2023), in which students worked in a laboratory for translation training, where they attended classes for a semester, and could use hard-copy dictionaries. The authors of the research argue that such setting allows to better explore the quality of the students' translations, as their decisions should be based on their knowledge, not online resources. However, one could

argue that when comparing quality and productivity of PE in contrast to purely translated texts, such setting does not allow participants to fully show their skills in both tasks. Being able to correctly use external resources and evaluate them based on knowledge is an inseparable skill of a good translator.

All in all, the best results come from a controlled environment (Translog-II) combined with access to external resources. Working in a CAT-tool and with external resources brought statistically more quality losses than gains. The PET environment did not bring any significant losses, and some gains were recorded. Using limited resources brought two contradictory, significant effects.

4.2.2.7 Chapter summary and conclusion

This chapter has focused on quality gains and losses of post-edited texts in comparison to translations produced from scratch, taking into account various factors such as language pairs, participant types, post-editing experience, translation systems, text types, and experiment settings. It has examined each factor and its connection to quality separately. The objective of this investigation was to deepen the understanding of how these diverse factors influence the quality of post-edited texts.

The results of all sections show the best quality for: EN to ZH translation, Google Translate, post-editing general texts, working in Translog-II with access to other resources, experienced translation students. The worse quality was reported in the following categories: students, no PE experience, using Google Translate and DeepL, the EN-SL language pair, working in a CAT-tool and using external resources, and translating technical texts. As evidenced by the results above, the same group of participants and the same MT systems can yield both the highest and lowest quality outcomes. Although the group of professional translators did not achieve the highest SQG scores, they demonstrated the most consistent results. Interestingly, access to various resources and the use of a CAT tool did not necessarily lead to higher quality outcomes. As anticipated, the quality of PEMT declined with the increasing technicality of the text.

5 Conclusion

5.1 Main findings and implications

This research examined the quality and productivity of PEMT in comparison to HT. The relationship between these two factors, as well as other aspects that may influence them, was investigated and presented in a clear manner. It aimed to discuss two research questions. Regarding RQ1, the analysis showed the following results:

1. Eighteen studies presented productivity gains without any corresponding changes in quality; however, sixteen of these studies did not utilize any form of quality assessment. The primary criterion for evaluation was the delivery of translations meeting publishable quality standards. The lowest recorded gain for this group was 1%, while the highest reached 370%. This observation suggests that in certain scenarios, the use of PE may not be particularly advantageous; however, it is likely that it will not result in any losses in quality.
2. Six studies reported significant and insignificant quality losses in case of productivity gains. Among these studies, the highest recorded productivity gain was 63%, while the lowest was almost 6%. In both cases, the quality loss was considered insignificant.
3. Seven cases demonstrated improvements in both productivity and quality, with the highest productivity gain approaching 77% and the least significant gain being just under 10%. The highest value noted a significant increase in quality, whereas the significance of the quality improvement for the lowest gain was not measured.
4. A small number of cases (five in total) showed productivity losses, with declines not exceeding -17%, while quality remained mostly unchanged. Four out of these five studies indicated productivity gains in the previously mentioned categories.

All in all, PEMT output primarily results in gains, with only a few instances of losses reported. While quality remains largely unchanged in most cases, it is important to highlight that quality assessment is often lacking in research studies, and the acceptable quality level is assumed. When quality is evaluated, the resulting productivity gains tend to be significantly lower.

In order to answer RQ2, sections 4.2.1 Productivity and analyzed factors and 4.2.2 Productivity and analyzed factors examined various factors influencing the PEMT productivity and quality:

A) Regarding productivity:

1. The following language pairs show a tendency to higher gains than other examined languages: EN-IT, EN-NL, EN-FR/FR-EN, EN-DE, HI-UR. On the other hand, EN-FI, EN-PL, EN-HI, EN-EL/EL-EN, EN-PT and EN-SV are more likely to show losses.
2. Professional translators exhibited the highest overall productivity gains, as evidenced by four studies reporting productivity levels exceeding 115%, all of which belong to this group. In contrast, translation students demonstrated lower productivity gains, with the four highest reported increases ranging from 53% to 77%. However, the translations produced by the second group, namely the students, showed fewer instances of losses; specifically, two studies indicated losses for students, whereas five cases reported losses across various scenarios for professional translators. This observation suggests that being a working translator does not necessarily guarantee optimal PE results.
3. PE-experience plays a big role. Three out of four overall highest gains, which exceed 130% and were reported by various studies, belong to this category. The absence of such experience does not necessarily lead to a decline in productivity; the highest three scores reported for inexperienced post-editors by different studies range from 84% to 119%. However, the results present a degree of ambiguity, as both experienced and inexperienced groups have also reported losses.
4. The use of commercial NMT systems, even in the absence of fine-tuning, yields the most consistent productivity outcomes. 23 out of 33 cases show gains between 30 and 77%. Three remaining ones are higher, and seven are lower. While fine-tuning an NMT system can lead to significant benefits—evidenced by achieving the highest productivity gains among all studies—it does not guarantee consistent results, as two cases reported the lowest scores. Furthermore, the effectiveness of machine translation systems integrated within CAT tools varies based on the language combination, with results ranging from - 9% for EN-SV to 148% for EN-FR.
5. When PE general texts, only one instance of loss was recorded, with the highest reported gain reaching 119%. 22 out of 33 cases reported gains of over 30%. PE specialized texts yielded impressive gains (the highest gains range from 130 to 370%). However, most losses (four), regarding different scenarios, can also be seen in this category. Research on literary texts reports consistent gains (between 18 and 36%) when using good quality NMT systems. Studies in which a collective gain was presented for a few various text types show quite different increase scores (from 5% to 71%).

6. Working in different experiment settings leads to various outcomes. Being able to translate and post-edit in a regular translation environment brings the highest gains (the highest scores were between 130 and 370%). However, this is not always the case. The results seem to be influenced by the language pair; Post-editing EN-SV resulted in losses of -9%, in contrast to scores between 4 to 148% for other language combinations in the same scenario. Limited resources do not cause the worst results; higher scores of just over 28% and 53% were recorded as well.

B) Regarding quality:

1. Post-editing texts in some of the language pairs tends to result in more gains than losses. All SQG were reported for an EN to ZH text. On the other hand, 3 out of 6 losses were recorded for this language pair as well.
2. PE has the greatest influence on the quality of texts produced by translation students. This means that it can be either beneficial (10 out of 14 gains were reported for this group) or detrimental (4 out of 6 losses) to the quality of their translation products. This group needs to be more cautious when performing the task due to the priming effect. Professional translators tend to mostly keep the same quality when PE or translating from scratch.
3. PE experience plays a crucial role; 13 out of 14 gains were recorded in this group. Therefore, this skill should be developed during university translation programs especially in the face of rapid translation technology developments.
4. NMT systems bring mixed results. In-house NMT provides the most stable outcome, whereas commercial NMT can either bring meaningful gains (10 out of 14 reported gains) or losses (3 out of 6 overall losses).
5. The quality of various text types differs. General text type profits the most from PE; seven gains were recorded. In general, greater text complexity tends to lead to more significant quality losses.
6. Experiment setting matters. However, just like in the case of productivity, working in a CAT-tool and having access to external resources does not guarantee gains. Two gains and three losses belong to this group. Working in Translog-II together with using other resources resulted in the most gains— nine out of fourteen.

My research confirms the conclusion on variety in translation speed made by Guerberof Arenas (2014) and Terribile (2024). Guerberof Arenas (2014) states that “it might be difficult to find a satisfactory solution to determine a “fair” price”, since the speed increase is rather individual

(Guerberof Arenas, 2014), and Terribile (2024) comes to a conclusions that the translation speed is highly variable. My research also corresponds with the remaining two findings of Terribile (2024); in most cases PE is faster than HT, and relying on average speed gains and losses can be deceiving. It has become evident that professional translation work is necessary to ensure high-quality texts, even for tasks that involve “just” post-editing. The need for modifications to the MT output proves that MT still requires human intervention.

5.2 Limitations and suggestions for future research

The main limitation of this research is the unequal number of studies exploring each of the factors. There was significantly more research available for some of the investigated aspects (for example eleven articles for the EN-ZH language pair and only one for the EN-UK). Having an even number of studies for each of the factors would ensure their comparability. The thesis revealed that gains were predominantly recorded. It is important to highlight that majority of these gains focused on language combinations that included English. For that reason, it cannot be assumed that PE brings the same results for other language pairs. Nevertheless, studies investigating other language pairs (DE-FR, DE-IT and HI-UR) did not report any productivity losses. Further exploration of post-editing in other languages is needed. It would be beneficial to the European market, where information about imported products often does not involve English, and it would provide useful insights into languages with higher complexity.

Secondly, regarding quality, in most cases it was not properly measured, which may lead to unreliable results, since a good quality is a subjective factor. That raises the question of whether PEMT can truly provide no changes to quality. While it is reasonable to expect that professional translators will deliver high-quality products, some experience from professional translators shows that not everyone who calls themselves a professional translator is able to deliver an objectively high-quality translation. Moreover, it cannot be expected that students will deliver translations at a “professional” level if they still have translation courses to complete. Of course, it may be the case, nevertheless it should not be assumed. Their perceived publishable level may be much lower than what is accepted in the industry.

Furthermore, due to the fact that MT quickly improves, more research need to be conducted, as the results from last year could differ from those published a couple of years ago.

Another aspect that requires more exploration are the criteria used for assessing quality and productivity. As shown in the analysis chapter, a wide range of categories and requirements

is applied, even if the main approach is the same (human evaluation). This makes it problematic to compare the results and does not provide clear answers. Therefore, establishing a common approach to analyze such studies or agreeing on a unified evaluation model would significantly advance research in the field of PE NMT. In terms of quality, it is essential to determine which aspects should matter to the industry and translators, ensuring that no sacrifices are made on either side. One potential solution would be to always report productivity and quality gains in percentages, since PE wages are usually lower by 30%. This would make it easier to determine whether a gain is sufficient when using a desired model, without imposing a single system. That would accelerate the process of analysis. On the other hand, when employing various models, we cannot be certain that the same phenomenon is being measured, nor whether the terms “quality” and “productivity” encompass the same factors in their definitions. Unifying models would simplify comparing results of different studies and enable future research to assess whether any changes occurred.

Additionally, the type of chosen translation environment for the experiment should be standardized as well. In the case of a controlled environment, it is impossible to assess what the situation looks like in a daily setting. Hypothetically, even if a proposed model does not fully measure the effort and quality, it would still enable sensible comparisons of results. Moreover, it would be possible to improve it with time.

Finally, there are many other aspects not included in this research that can influence the quality and productivity of PE texts, such as attitude or being a native or non-native speaker of the target language. Since this research investigated six broad factors, it was not possible to explore additional factors. Nonetheless, it is crucial to identify which of these factors hold the greatest significance. Productivity is a complex construct that cannot be fully understood by examining only its temporal dimension. It is important to recognize that true productivity results from a combination of various factors, each encompassing different types of effort.

MT proved to be useful to translators. It should not be seen as a threat to our profession, but rather as a practical tool that makes our work easier. Nevertheless, when post-editing, one needs to be cautious, as there are cases that report losses both in quality and productivity, which could have serious consequences and make our work unprofitable. Post-edited texts are reported to have lower lexical variety (Torral, 2019, p. 279), which could lead to impoverishment of a language. This is a view shared by Farrell (2018), who proved that MT-output may lack the variety and inventiveness of HT (Farrell, 2018, p. 50). Another problem reported is an increased percentage of interference regarding sentence length and grammatical structures (Torral, 2019,

p. 279). The reason named above could lead to the disappearance of grammatical structures typical for a certain language. Nevertheless, it cannot be denied that translators can strongly profit from PE NMT; therefore, this topic should be explored more.

References

- Ahmad, R. et al., 2018. *Transzaar: Empowers Human Translators*. Melbourne, VIC, Australia, IEEE.
- Ahsan, A., Mujadia, V. & Sharma, D. M., 2021. *Assessing Post-editing Effort in the English-Hindi Direction*. Silchar, National Institute of Technology Silchar, pp. 44-53.
- Álvarez-Vidal, S., Oliver, A. & Badia, T., 2020. Post-editing for Professional Translators: Cheer or Fear?. *Revista Tradumàtica. Tecnologies de la Traducció*, 18, pp. 49-69.
- Anon., 2023. *Machine Translate*. [Online]
Available at:
<https://machinetranslate.org/ter#:~:text=Translation%20Error%20Rate&text=TER%20is%20defined%20as%20the,average%20length%20of%20the%20references.&text=TER%20corresponds%20to%20post%20editing%20effort>.
[Accessed 07 August 2024].
- Bassnett, S. & Lefevere, A., 1998. *Constructing Cultures. Essays on Literary Translation*. Clevedon: Multilingual Matters.
- Béchara, H., Orasan, C., Escartín, C. P. & Lowe, W., 2021. The Role of Machine Translation Quality Estimation in the Post-Editing Workflow. *Informatics*.
- Biel, Ł., 2021. Postedycja tłumaczeń maszynowych. *Lingua Legis nr 29*, pp. 11-34.
- Bogucki, Ł., 2009. *Tłumaczenie wspomagane komputerowo*. Warsaw: Wydawnictwo Naukowe PWN.
- Carl, M. et al., 2011. The Process of Post-Editing: a Pilot Study. *Copenhagen Studies in Language*, Volume 41, pp. 131-142.
- Carl, M. & Schaeffer, M., 2018. *Researching Cognitive Processes of Translation*. Singapore: Springer Nature Singapore.
- Choudhury, R. & McConnell, B., 2013. *Translation Technology Landscape Report*, Amsterdam: Translation Automation Users Society TAUS.
- Correia, G. M. & Martins, A. F. T., 2019. A Simple and Effective Approach to Automatic Post-Editing with Transfer Learning. In: *Proceedings of the 57th Annual Meeting of the*

- Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 3050–3056.
- Cui, Y., Liu, X. & Cheng, Y., 2023. A Comparative Study on the Effort of Human Translation and Post-Editing in Relation to Text Types: An Eye-Tracking and Key-Logging Experiment. *Sage Open*, pp. 1-15.
- Daems, J., Vandepitte, S., Hartsuiker, R. J. & Macken, L., 2017. Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. *Meta Journal des traducteurs Translators' Journal*, pp. 245-270.
- Dede, V., 2022. *Temporal and Technical Effort in Post-editing compared to editing and translating from scratch*. Master's Thesis, Ankara: Hacettepe University Graduate School of Social Sciences.
- Depraetere, I., 2011. *Perspectives on translation quality*. Berlin: de Gruyter Mouton.
- Cambridge Dictionary, 2023. [Online]
Available at: <https://dictionary.cambridge.org/dictionary/english-german/productivity>
[Accessed 01 July 2023].
- do Carmo, F., 2020. “Time is money” and the value of translation. *Translation Spaces* 9 (1), pp. 35-57.
- do Carmo, F. et al., 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, pp. 101-143.
- Escartin, C. P. & Arcedillo, M., 2015. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation. *Proceedings of MT*, 1(15), pp. 131-144.
- Farrell, M., 2018. Machine Translation Markers in Post-Edited Machine Translation Output. *Proceedings of the 40th Conference Translating and the Computer*, pp. 50-59.
- Farrell, M., 2023. Current evidence of post-editeese: differences between post-edited neural machine translation output and human translation revealed through human evaluation. *Proceedings of the International Conference HiT-IT*, pp. 52-63.
- Forcada, M. L., 2017. Making sense of neural machine translation. *Translation Spaces*, pp. 291-209.

- Fulford, H., 2002. Freelance translators and machine translation: An investigation of perceptions, uptake, experience and training needs. *Paper presented at the 6th annual conference of the European Association for Machine Translation Workshop on Teaching Machine Translation*, pp. 117-122.
- Görög, A., 2015a. *TAUS the language data network*. [Online] Available at: <https://www.taus.net/resources/blog/translation-productivity-revisited> [Accessed 03 07 2023].
- Görög, A., 2015b. *The TAUS Efficiency Score. Introducing a new score for measuring productivity*. Dublin, Paper presented at the TAUS Quality Evaluation Summit, 28 May 2015.
- Guerberof Arenas, A., 2014. Correlations between productivity and quality when post-editing. *Machine Translation*, pp. 165-186.
- Guerberof, A., 2009. *Productivity and quality in MT post-editing*, Ottawa: MTSummit.
- Hao, Y., Hu, K. & Pym, A., 2024. Who's Afraid of Literary Post-editing? Performances and Reflections of Student Translators. *New Advances in Translation and Interpreting Technology: Theories, Applications and Training*, pp. 263-282.
- House, J., 1997. *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Nar Verlag Tübingen.
- Huang, J. & Wang, J., 2022. Post-editing machine translated subtitles: examining the effects of non-verbal input on student translators' effort. *Perspectives*, pp. 1-21.
- Hutchins, W. J., 1994. Machine Translation: History and General Principles. In: R. E. Asher, ed. *The encyclopedia of languages and linguistics*. Oxford: Pergamon, pp. 2322-2332.
- ISO:18587, 2017. *ISO 18587*
- Jakobsen, A. L. & Mesa-Lao, B., 2017. *Translation in Transition. Between Cognition, Computing and Technology*. Amsterdam: John Benjamins Publishing Company.
- Jiang, Z. & Zhang, Z., 2024. *Convergences and Divergences between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation*, Shanghai: Shanghai Jiao Tong University.

- Jia, Y., Carl, M. & Wang, X., 2019a. Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation Vol. 33. No. 1–2.*, pp. 9-29.
- Jia, Y. & Si, L., 2022. Post-Editing Metaphorical Expressions: Productivity, Quality, and Strategies. *The Journal of Foreign Languages and Cultures Vol. 6 Nr. 2*, pp. 28-43.
- Jia, Y. & Sun, S., 2022. Man or machine? Comparing the difficulty of human translation versus neural machine translation post-editing. *Perspectives VOL. 31, NO. 5*, pp. 950-968.
- Jia, Y., Wang, X. & Carl, M., 2019b. How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *The Journal of Specialised Translation, Issue 31*, pp. 60-86.
- Killman, J. & Rodríguez-Castro, M., 2022. Post-editing vs. translating in the legal context: Quality and time effects from English to Spanish. *Revista de Llengua i Dret, Journal of Language and Law*, 78, pp. 57-72.
- Koby, G. S. et al., 2014. Defining Translation Quality. *Revista Tradumàtica: tecnologies de la traducció*, December, 12(1578-7559), pp. 413-420.
- Koehn, P., 2020. *Neural machine translation*. Cambridge: Cambridge University Press.
- Kolena, n.d. *Testing with Kolena*. [Online]
Available at: <https://docs.kolena.com/metrics/bleu/>
[Accessed 07 August 2024].
- Koller, W., 1995. The concept of equivalence and the object of translation studies. *Target*, 7(2), p. 191–222.
- Kuzman, T., Vintar, Š. & Arčan, M., 2019. Neural Machine Translation of Literary Texts from English to Slovene. *The Qualities of Literary Machine Translation*, pp. 19-23.
- Läubli, S. et al., 2019. Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 267-272.
- Lavie, A., 2020. *Unbabel*. [Online]
Available at: <https://unbabel.com/why-we-built-comet-a-new-framework-and-metric-for-automated-machine-translation-evaluation/>
[Accessed 7 August 2024].

- Lavie, A. & Denkowski, M. J., 2009. The METEOR metric for automatic evaluation of Machine Translation. *Machine Translation*, Volume 23, pp. 105-115.
- Lelner, Z., 2022. *MemoQ Blog*. [Online]
Available at: <https://blog.memoq.com/machine-translation-vs.-machine-translation-post-editing-which-one-to-use-and-when>
- Li, J. & Way, A., 2021. *Machine Translation*. 1st ed. New York City: Springer Publisher.
- Lommel, A. & Pielmeier, H., 2021. *Machine Translation Use at LSPs: Insights on How Language Service Providers' Use of MT Is Evolving*, San Francisco: Common Sense Advisory Research.
- Macken, L., Prou, Daniel & Tezcan, A., 2020. Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process (European Commission Directorate-General for Translation).
- Martín, J. A. A. & Serra, A. C., 2014. Integration of a machine translation system into the editorial process flow of a daily newspaper. *Procesamiento del Lenguaje Natural*, Issue 53, pp. 193-196.
- Mathur, N., Baldwin, T. & Cohn, T., 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4984–4997.
- McEnery, T. & Hardie, A., 2011. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- Moorkens, J. et al., 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29 (3), p. 267–284.
- Moorkens, J., Toral, A., Castilho, S. & Way, A., 2018. Translators' Perceptions of Literary Post-editing using Statistical and Neural Machine Translation. *Translation Spaces issue 7:2*.
- Murgolo, E., 2022. Productivity Evaluation in MT Post-Editing and Fuzzy Matches. *New Trends in Translation and Technology*, pp. 1-8.
- Negri, M. T., Bertoldi, N. & Federico, M., 2018. Online neural automatic post-editing for neural machine translation. *Proceedings of the 11th Italian conference computational linguistics*, Volume 2253.
- Nida, E. & Taber, C. R., 1969. *The Theory and Practice of Translation*. Leiden: Brill.

- Nitzke, J. & Hansen-Schirra, S., 2021. *A short guide to post-editing*. Berlin: Language Science Press.
- Nizam, M., 2024. *Neural Networks: From Biological to Artificial*. [Online]
Available at: <https://www.linkedin.com/pulse/neural-networks-from-biological-artificial-muhammed-nizam-pcxwf/>
[Accessed 22 March 2024].
- O'Brien, S., 2002. *Teaching Post-editing: A Proposal for Course Content*, Dublin City: Dublin City Libraries.
- O'Brien, S., 2021. Post-editing. In: Y. Gambier & L. van Doorslaer, eds. *Handbook of Translation Studies*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 177-185.
- Offer, T., 2024. *Tomedes*. [Online]
Available at: <https://www.tomedes.com/translator-hub/mtpe>
[Accessed 10 August 2024].
- Pavlíková, D., 2022. Comparing Human Translation and Machine Translation Post-Editing. *L10N Journal*. 1(1), pp. 66-105.
- Pierce, G., 2018. *Introducing Translational Studies*. Amsterdam: Elsevier.
- Plitt, M. & Masselot, F., 2010. Productivity test of statistical machinetranslation post-editing in a typical localisation context. *Prague Bull. Math.Linguist.* 93, pp. 7-16.
- Priddy, K. L. & Keller, P. E., 2005. *Artificial Neural Networks*. Bellingham, Washington: SPIE Press.
- Rajarman, V., 2024. *Anecdotes from the history of modern computing*. Delhi: PHI.
- Ramos, F. P., 2021. *Institutional Translation and Interpreting Assessing Practices and Managing for Quality*. New York and London: Routledge, Taylor & Francis Group.
- Rothwell, A. et al., 2023. *Translation Tools and Technologies*. Milton Park: Taylor & Francis.
- Ruopp, A., 2015. Industry Shared Metrics with the TAUS Dynamic Quality Dashboard and API. *Proceedings of MT Summit XV, vol. 2: MT Users' Track*, 30-3 October-November, pp. 163-191.

- Sabry, F., 2023. *Perceptrons - Fundamentals and Applications for The Neural Building Block*, London: One Billion Knowledgeable.
- Sai, A. B. et al., 2022. *IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages*, Stroudsburg: Association for Computational Linguistics.
- Sánchez-Gijón, P., Moorkens, J. & Way, A., 2019. Post-Editing Neural Machine Translation vs. Translation Memory Segments. *Machine Translation*, pp. 1-29.
- Sarti, G., Bisazza, A., Guerberofo-Arenas, A. & Toral, A., 2022. DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages.
- Sismat, M. A. H., 2019. Inverse Translation Quality: A Comparative Analysis between Human Translation and Post-editing. | *Journal of Arabic Linguistics and Literature*, pp. 91-105.
- Sivanandam, S. & Paulraj, M., 2009. *Introduction to Artificial Neural Networks*. Delhi: Vikas Publishing House Pvt Limited.
- Somers, H., 1999. Review Article: Example-based Machine Translation. *Machine Translation*, Volume 14, pp. 113-157.
- Sosoni, V., O'Shea, J. & Stasimioti, J. M., 2022. Translating Law: A Comparison of Human and Post-edited Translations from Greek to English. *Revista de Llengua i Dret, Journal of Language and Law*, 78, pp. 93-108.
- Stasimioti, M. & Sosoni, V., 2020. Translation vs Post-editing of NMT Output: Measuring effort in the English-Greek language pair. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, pp. 109-124.
- Stasimioti, M. & Sosoni, V., 2021. Investigating post-editing: A mixed-methods study with experienced and novice translators in the English-Greek language pair. *Translation, interpreting, cognition: The way out of the box*, pp. 79-104.
- Stefaniak, K., 2020. Evaluating the usefulness of neural machine translation for the Polish translators in the European Commission (DGT). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 263–269.
- TAUS, 2024. *TAUS the language data network*. [Online]
Available at: <https://www.taus.net/>
[Accessed 8 August 2024].

- Terribile, S., 2024. *Productivity in the Post-Editing of Neural Machine Translation: A Mixed-Methods Analysis of Speed and Edits at Toppan Digital Language*, Manchester: Doctoral Thesis.
- Toral, A., 2019. Post-editeese: an Exacerbated Translationese. *Proceedings of MT Summit XVII, volume 1*, pp. 273-281.
- Toral, A., Wieling, M. & Way, A., 2018. Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities*, Volume 5, pp. 1-11.
- Toury, G., 1980. *In Search of a Theory of Translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics.
- Valli, P., 2015. The TAUS Quality Dashboard. *Proceedings of the 37th Conference Translating and the Computer*, p. 127–136.
- Vasconcellos, M., 1986. Post-editing on-screen: machine translation. *Translating and the Computer*, 13-14 November, pp. 133-146.
- Walker, C. & Federici, F. M., 2018. *Eye Tracking and Multidisciplinary Studies on Translation*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Wang, J., 2022. Measuring Student Translators' Cognitive Effort with Pauses: A Comparative Analysis of Human Translation and MT. *Communication and Linguistics Studies*. Vol. 8, No. 4, pp. 80-84.
- Wang, X., Wang, T., Mart, R. M. & Jia, Y., 2021. *Investigating usability in postediting neural machine translation: Evidence from translation*, Budapest: Akademiai Kiado.
- Williams, M., 2004. *Translation Quality Assessment: An Argumentation-Centred Approach*. Ottawa: University of Ottawa Press.
- Winther Balling, L., Carl, . & O'Brian, S., 2014. *Post-editing of Machine Translation: Processes and Applications*. Cambridge: Cambridge Scholars Publisher.
- Yamada, M., 2014. Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation*.
- Yamada, M., 2019. The impact of Google Neural Machine Translation on Post-editing by student translators. *The Journal of Specialised Translation Issue 31*, pp. 87-105.

- Yang, Y., Liu, R., Qian, X. & Ni, J., 2023. Performance and perception: machine translation post-editing in Chinese-English news translation by novice translators. *Humanities and Social Sciences Communications*, pp. 1-9.
- Zavala-Rojas, D. et al., 2024. Using Machine Translation and Post-Editing in the TRAPD Approach: Effects on the Quality of Translated Survey Texts. *Public Opinion Quarterly Vol 88 No 1*, p. 123–148.
- Zhang, J. & Qian, H., 2023. The impact of machine translation on the translation quality of undergraduate translation students. *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pp. 99-108.
- Zuber, R., 1968. *The belles infideles and the formation fo classic taste*. Paris: Colin.

Appendices

Appendix I Table with all productivity and quality gains/losses

RQ1: What is the relationship between translators' productivity when post-editing machine translation and the translated text's quality?			
Article	Measure	Productivity	Quality
(Ahsan, Mujadia, & Sharma, 2021)	a) source words per hour b) human evaluation	G and L average 63% time savings =(150,2% to -2,6% in time savings)	G -very slight (p-value=0.08) I
(Álvarez-Vidal & Oliver, 2020)	a) source words per hour b) not measured	G average 53.14% (between 33.29% to a 81.01%)	E not measured
(Béchara, Orasan, Escartin i Lowe, 2021)	a) scores of the translators' sentences, comparing them to the post-edited reference provided by Autodesk b) Second per source word	a) G 24% [3,8 s per word vs 2,9 s per word] →no MTQE b) G 37% [3,8 s per word vs 2,4 s per word] → MTQE	G Approximate value No QE 5% QE10%
(Cui, Liu, & Cheng, 2023)	a) seconds per target word b) human assessors: 100-mark system according to the following criteria: accuracy of meaning, fluency of expression, and reproduction of stylistic characteristics. + self-rating	G advertising text: 38,4% (Post-editing ME= 2.882 HT: ME= 4.757) News text: 36,3% (HT= ME= 3.778, PE= E= 2.406) legal text: 53,6% (HT = ME= 4.976, PE=M E= 2.308) poetic text: 23,4 % (HT= ME= 6.467, PE= ME= 4.956.)	G, L Advertising text: (HT = ME= 84.667, PE= ME= 90.333) 6,7% News text: 3,8% (HT= ME= 85, PE= ME= 88.333) legal text: -0,09% (ME= 86.667, ME= 86) poetic text: 7,3% (ME= 77.667, ME= 83.333)
(Dede, 2022)	a) Seconds per segment b) Not measured	G translation from scratch mean 50.69 Google mean 30.29 Average= 40,2% Google 1.0 - 42,1% Google high - 37% Google low - 41,5%	E not measured
(Farrell, 2018)	a) Minutes per text b) number of errors: acceptable, debatable, mistranslation	G Venice text 5.94%, Verona text 14.94%	L (2.04 vs 2.25 errors per translator) 10,2%
(Farrell, 2023)	a) the translation time minus the post-editing time, divided by the translation time.	G Text about Ukraine 47.99% Text about Slovakia 39.50%	E not measured, guidelines for quality, publishable quality
(Huang & Wang, 2022)	a) Seconds per segment	G 50% (8,75 = average PEMT 17,5 =average HT)	E not measured
	b) Guidelines for quality		
(Jia & Si, 2022)	a) Seconds per text quality. b) Customized error types developed by the Multidimensional Quality Metrics	G (in seconds 9.29 vs 7.23) 32,2%	G. There were 69 errors in the PE and 88 in the HT groups' output and the difference was statistically significant (t=2.24; p<0.05; df=28; ES=0.83) 21,6%
(Jia & Sun, 2022) sprawdź czy były straty	a) Words per second b) Number of errors, fluency and accuracy. Human assessor, terms of accuracy (mistranslation, addition, and omission) and fluency (word form, cohesion, word order, punctuation, unintelligible, and spelling), based on the customized error categories of the Multidimensional Quality Metrics.	For STh (source text with high complexity) 1) HT (M = 0.13) vs PEMTH (M = 0.23) vs PEMTL (M = 0.12) → 1a) high quality machine – 76,9% 1b) Low quality machine – -7,7% 1. ST1 (source text with low complexity) 2) HT (M=0.14) ,PEMTH (M=0.17) PEMTL (M= 0.16) → 2a) high q. m – 21,4% 2b) lqm 14,3% G High quality google translate, systna low	G Source Text of high complexity: a) accuracy significantly higher, H24%, L38% b) fluency: fluency, higher but not significantly Accuracy: 2. For STh (source text with high complexity) 1) 1a) high quality machine – 1b) Low quality machine– 33,5% 3. ST1 (source text with low complexity) 39,2% 2) HT () ,PEMTH () PEMTL () → 2a) high q. m 20,9% 2b) lqm 7%
(Jia, Carl, & Wang, 2019) 2019a	a) milliseconds per segment b) Quality: guidelines TAUS's Dynamic Quality Evaluation Framework. Not measured!	G Text1 55% Text2 72% (Text 2 is of higher complexity than Text 1 as indicated by human rating of text difficulty, readability level, word frequency and the number of non-literal expressions (idiom, metonyms and metaphors). 1. Text1 4000 do 9000 →55% faster 2. Text 2 2280 vs 9800 →72% faster	E not measured
(Jia, Wang, & Carl, 2019)b	A) time per word B) Human evaluators	G (significant for domain specific texts) domain-specific trans 11000 p. 7600 Domain specific: 31%	E ..
		general trans 10500 vs p10150 3% 3%	General:

(Kuzman, Vintar, & Arčan, 2019)	a)translation time b)not measured. Only the raw output of MT was rated	G but; 1,6% Google Neural Machine Translation (GNMT) (post.editing =10,1 average human13,0 =28,7% IG • model, trained on out-of-domain data and retrained on the corpus SPOOK (model 'SPOOK') (post-editing= 52,4 vs human54,5) 4% • model, trained on out-of-domain data and retrained on the novel Practice Makes Perfect (model 'Novel') (post52,1 vs human44,4) loss -16,8%	E not measured
(Läubli, et al., 2019)	A)Words per hour b) 1-6 scale	G 59,74%	E 4.50/450, no loss
(Läubli, et al., 2019)	A)Words per hour B)1-6 scale	G 9,26%	G slight gain 4,75, --4,88
(Macken, Prou, Daniel, & Tezcan, 2020)	A)Seconds per word B) Not measured	G14% average (between 11-21%) BUT: When provided with such perfect MT suggestions = 66% faster, compared with translating from scratch. WHEN NO EDITING IS NEEDED	E not measured
(Moorkens, Toral, Castilho, & Way, 2018)	Second per source text characters	27,25 % approximate value	E not measured
(Murgolo, 2022)	a)Words per hour b) not measured	Mechanical patents: 370% trained engine Generic engine 324,85% Life Science Patents: 367% trained 318% general TLS and electronics patents: 252% trained 215% general *this group has the least experience in translation and post-editing	E not measured
(Sarti, Bisazza, Guerberof-Arenas, & Toral, 2022)1	a) Seconds per word b) Not measured	a) G 84% Google Translate b) G 10 % MP-BART 50	E not measured
(Sarti, Bisazza, Guerberof-Arenas, & Toral, 2022)2	B) Seconds per word C) Not measured	a) G 119% b) G 61 %	E not measured
(Sarti, Bisazza, Guerberof-Arenas, & Toral, 2022)3	a) Seconds per word b) Not measured	a) G 96% b) G 95 %	E not measured
(Sarti, Bisazza, Guerberof-Arenas, & Toral, 2022)4	a) Seconds per word b) Not measured	a) G 34 % b) G 12 %	E not measured
(Sarti, Bisazza, Guerberof-Arenas, & Toral, 2022)5	a) Seconds per word b) Not measured	a) G 71 % b) G 14 %	E not measured
(Sarti, Bisazza, Guerberof-Arenas, & Toral, 2022)6 17	a) Seconds per word b) Not measured	a) G 32 % b) G 23 %	E not measured
(Sosoni, O'Shea, & Stasimioti, 2022)	a)Minutes per text b) MQM-DQF error typology, modified	L 9% (38 minutes on average to translate the texts from scratch, and 41.6 minutes on average to post-edit the MT output) ONLY TWO TRANSLATORS Translator A needed 23 minutes for translation from scratch and 23.2 minutes for PE, while Translator B needed 53 minutes and 60 minutes, respectively.	G 21 vs 2o errors on average . 5%
(Stasimioti & Sosoni, 2020)	a)Minutes per text b) not measured	G19.8% (average) worth noting: Translations general: 9.86 (not specified which group, all together) PE with experience: 7.07 = 28,3% PE no experience: 10.42 = -5,7% (loss!) L	E not measured
(Stefaniak, 2020)	a) words per seconds (minutes per page) b)not measured	G 17% per page (0.20 vs 0.32 words/seconds) productivity gain = (time to edit TM matches + time to translate from scratch) – (time to edit TM matches + time to postedit MT)	
(Terribile, 2024)	A)words per hour B) not measured	G, L EN—DA 29%, WO 30% (without outliers) EN-DE 63% 74% EN-ES 67% 67% EN-FI 6% 4% EN-FR 130% 148% EN-IT 89% 90% EN -NL 62% 69% EN-PL 18% -4% EN-PT 5% 8% EN-SV -7%, -9%	E not measured
		FR-EN 85% 86%	

(Torral, Wieling, & Way, 2018)	Words per hour	G 36% (however varies from 1-2% to over 80% for different translators)	E not measured
(Wang J. , 2022)	A)time per words B) not measured	G 44,7% 2758 ms and for human translation is 4223ms	E not measured
(Wang, Wang, Mart., & Jia, 2021)	A)words per second b)Human reviewers using Taus's Dynamic Quality Evaluation Framework. Fluency and adequacy, plus reference translations	G growth (P < 0.01) approximate value 53,9% (regular unaided translation)	G ... (P < 0.01), Wilcoxon test Adequacy> 54% scores MTPE (M 5 3.173) and RUT (M 5 2.051) Fluency – 27,3% (M 5 3.214) and RUT (M 5 2.337)
(Yang, Liu, Qian, & Ni, 2023)	A)Minutes per text b)Human reviewers, rating guidelines	G Z = -2.53, p = 0.01 .. (20vs 30 minutes) 43,4%	G.. less than 1% [mean 84,58]
(Zhang & Qian, 2023)	a) time per text b) TAUS Harmonised DQF-MQM Error Typology (accuracy, fluency, style, variety)	G 10,2% from-scratch translations (M = 70.50) PEMT-assisted translations (M = 63.30)	L total error count = 11,2% from-scratch translations (M = 22.30) PEMT (M = 24.80) only type I accuracy errors 24,3% HT (M = 14.4) were significantly lower than PEMT-assisted translations (M = 17.90)
(Pavliková, 2022)	a)time per task b) TAUS quality assessment template. <u>thirty three</u> types of errors are included; they cover morphological, lexical, stylistic, syntactic, terminological, formatting, localization, and cultural issues.	G 29,4% 25.9 vs 36.8min 29,4%	. translation from scratch 7.8 vs 10 post-editing*average error count)
(Ahmad, et al., 2018)	A) Words per day B) Not measured	ENG-HI novice 96,2% G EN-HI beginner 77% Medium EN-HI 66% HI-UR expert 161% EN-HI expert 171%	E not measured
(Hao, Hu, & Pym, 2024)	a) Seconds per word b) Human reviewers	18%	3,33 vs 2,58 errors L (p < .001)+ t (p = .873)

Appendix II Table with all factors influencing productivity and quality gains/losses

Research question 2: Do PEMT productivity and quality differ according to domains, language pairs, and types of MT-post-editing?

Article	Domain	LAN	PE	Setting/tools	Engine	Participants	Exp-PE
1. (Ahsan, Mujadia, & Sharma, 2021)	News articles	EN-HI	Full	~~Indictranslate~~	Trained ~~MultiIndicMT~~	Professional Translators	Yes 70%/No 30%
2. (Álvarez-Vidal & Oliver, 2020)	General texts	ES-EN	Full	PET (other resources - not mentioned?)	~~DeepL~~	Professional Translators	No/Yes 42% some training
3. (~~Béchara, Orasan, Escartín i Lowe, 2021~~)	General sentences 260 sentences ~~rom~~ the Autodesk Post-editing Parallel Corpus	EN-ES	Full	PET = CAT Tool	In-house MT engine from Autodesk	Professional Translators	Yes
4. (Cui, Liu, & Cheng, 2023)	Advertising text, News text (general text), Legal text, literary text	ENG-ZH	Full	~~Translog-II~~ , electronic dictionary	Google Translate	Translation Students	Yes
5. (Dede, 2022)	General text Regular sentences	EN-TR	Full	SDL Trados 2017. ~~Quality~~	Google Translate	Professional Translators	75% No, the rest had some
6. (Farrell, 2018)	GT Wikipedia (Venice, Verona texts)	EN-IT	~~Full~~	Dictionaries, web resources etc. Microsoft Word	Google Translate	Translation Students	Yes
7. (Farrell, 2023)	~~GT Texts~~ from Wikipedia	EN-IT	~~Full~~	All external resources, working in a Word document	Pay-for version of ~~DeepL~~	Translation Students	Yes

8. (Huang & Wang, 2022)	Subtitles law mixed Drama documentary	EN-ZH	Full	~~Translog~~ (resources – not mentioned?)	Google Translate	Master Translation Students	No professional training or experience
9. (Jia & Sun, 2022)	GT News texts	EN-ZH	~~Full~~	EyeLink 1000 plus (SR Research) and their keystrokes with ~~Translog-II~~ .	Google Translate and SYSTRAN (the best and worst)	Translation Students	Yes PE training before the experiment
10. (Jia, Carl, & Wang, 2019) A	GT Newspaper texts *General texts	EN-ZH	Full	Online dictionaries or other external resources. ~~Translog-II~~	Google Translate	Translation Students	No
11. (Jia, Wang, & Carl, 2019) B	Domain specific: patient information leaflet, and a dishwasher manual General language texts: promotional brochures for Coca-Cola and Starbucks	EN-ZH	~~Full~~	External internet sources. All sources allowed. ~~Translog-II~~	Google Translate	Translation Students	No
12. (Kuzman, Vintar, & Arčan, 2019)	Literary texts,	EN-SL	~~Full~~	Post-Editing Tool (PET) interface (Aziz et al., 2012), a CAT tool	Google Translate, bespoke NMT models, one tailored to novels	Translation Students	Yes (83%)

13. (Läubli, et al., 2019)	Banking and finance	GE-FR	Full	Domain-specific translation memories and term bases, any online services	Domain specific trained (WMT17-style bi-RNN),	Professional translators	Yes
14. (Läubli, et al., 2019)	Banking and finance	GE-IT	Full	Domain-specific translation memories and ~~termbases~~ , online services	~~Trained~~ (WMT17-style bi-RNN	Professional translators	Yes
15. (Macken, Prou, Daniel, & Tezcan, 2020)	DGT- legislative documents and external communication documents	EN-FI	Full	Translation memories and ~~termbases~~	Directorate-General for Translation (DGT) – MT in Trados Studio	Professional translators	Yes
16. (Moorkens, Toral, Castilho, & Way, 2018)	Novel	EN-CA	~~Full~~	PET (Post-Editing Tool) interface	System tailored for literal translation	Professional Translators	Yes and No
17. (Murgolo, 2022)	patents in the mechanical, TLC and electronics, life sciences	EN-IT	Full	CAT tool with data base	NMT based on ~~OpenNMT-Lua~~ , trained with specific in-domain	Professional Translators	Yes
18. (Sarti, Bisazza, Guerbero-Arenas, & Toral, 2022) 2	GT ~~Sentences English~~ Wikipedia, mixed topics and domains	EN-NL	Full	PET, controlled set-up	Google Translate ~~mBART~~	Professional translators	Yes
19. (Sarti, Bisazza, Guerbero-Arenas, & Toral, 2022) 3	GT ~~Sentences English~~ Wikipedia, mixed topics and domains	EN-IT	Full	PET, controlled set-up	Google Translate ~~mBART~~	Professional translators	Yes

20. (Sarti, Bisazza, Guerbero-Arenas, & Toral, 2022) 4	<u>GT</u> <u>Sentences</u> , English Wikipedia, mixed topics and domains	EN-TR	Full	PET, controlled set-up	Google Translate <u>mBART</u>	Professional translators	Yes
21. (Sarti, Bisazza, Guerbero-Arenas, & Toral, 2022) 5	<u>GT</u> <u>Sentences</u> , English Wikipedia, mixed topics and domains	EN-UK	Full	PET, controlled set-up	Google Translate <u>mBART</u>	Professional translators	Yes
22. (Sarti, Bisazza, Guerbero-Arenas, & Toral, 2022) 6	<u>GT</u> <u>Sentences</u> , English Wikipedia, mixed topics and domains	EN-VI	Full	PET, controlled set-up	Google Translate <u>mBART</u>	Professional translators	Yes
23. (Sarti, Bisazza, Guerbero-Arenas, & Toral, 2022) 1	<u>GT</u> <u>Sentences</u> , English Wikipedia, mixed topics and domains	EN-AR	Full	PET, controlled set-up	Google Translate <u>mBART</u>	Professional translators	Yes
24. (Sosoni, O'Shea, & Stasimioti, 2022)	Legal texts	EL-EN	Full	They could use resources such as paper and online dictionaries and databases like IATE, but they were told not to use any CAT tools, translation memories or MT systems. <u>Working in a word document</u>	Domain trained	Professional Translators	No/Yes – training session
25. (Stasimioti & Sosoni, 2020)	educational texts OER Commons	EN-EL (Greek)	Full	<u>NO</u> External resources (dictionaries, Internet, etc)	Google Translate	Professional translators	Yes and <u>No</u>

26. (Stefaniak, 2020)	¹ Legislative and non-legislative mixed	EN-PL	Full	translation memories (<u>Euramis</u>) and a terminology database (IATE). <u>Regural</u> setup	Directorate General for Translation (DGT) MT in a cat tool <u>eTranslation</u>	Professional translators	Yes
27. (Terribile, 2024) (9 language combinations)	Marketing texts, <u>technical</u> marketing	EN-DA EN>D E EN>E S EN>FI R EN>IT EN>N L EN>P L EN>P T EN>P S	Full	<u>MemoQ</u> , Translation Memory	Translate Media - eCAT environment, a <u>customised</u> version of MT in <u>MemoQ</u>	Professional translators	Yes

		FR-EN					
28. (Toral, Wieling, & Way, 2018)	Novel	EN-CA	Full	PET	System tailored for literal translation	Professional Translators	?
29. (Wang J. , 2022)	data in TPR-DB	EN-ZH	Full	Translog II, resources	Baidu Translate	English Students (basic translation courses)	?
30. (Wang, Wang, Mart., & Jia, 2021)	General texts	EN-ZH	Full	No access to the Internet or other resources, controlled set-up. Translog-II	Google Translate	Translation trainees	No
31. (Jia & Si, 2022)	Texts with metaphorical expressions	EN-ZH	Full	Translog-II, access to the Internet and other external resources	Google Translate	Translation Students	Yes-training
32. (Yang, Liu, Qian, & Ni, 2023)	GT News texts	ZH-EN	Full	Online resources, everything allowed	Google Translate	Translation learners	Yes
33. (Zhang & Qian, 2023)	Specialized e texts were selected from the Accreditation Test for Translators and Interpreters in China (CATTI).	EN-CH	Full	a laboratory for translator training in which the students had attended classes for one semester, hard-copy dictionaries	Not mentioned	Translation Students	No, +little experience in translating
34. (Pavliková, 2022)	Specialized marketing text (American brochure on com posting; specifically, it was a manual on how to	EN-SL	Full	Memsources, terminology bank Yes	Google Translate	Translation Students	No *guidelines

	build a home vermicompost. It was a specialized marketing text)						
--	---	--	--	--	--	--	--

	marketing text)						
35. (Ahmad, et al., 2018)	Hindi news media, English exam books, semi-legal government documents, e-learning websites, documentary subtitles. Mixed, GT	EN-HI HI-UR	Full	Transzarr – an AI aided toll with CAT functionality	Transzarr	Professionals	Yes/ two weeks of experiment
29							
(Hao, Hu, & Pym, 2024)	Literary translation – texts by Agatha Christie	EN-ZH	Full	External resources	DeepL	Translation Students	Yes (no formal training)

Appendix III Post-editing and human translation productivity (Jia, et al., 2019a)

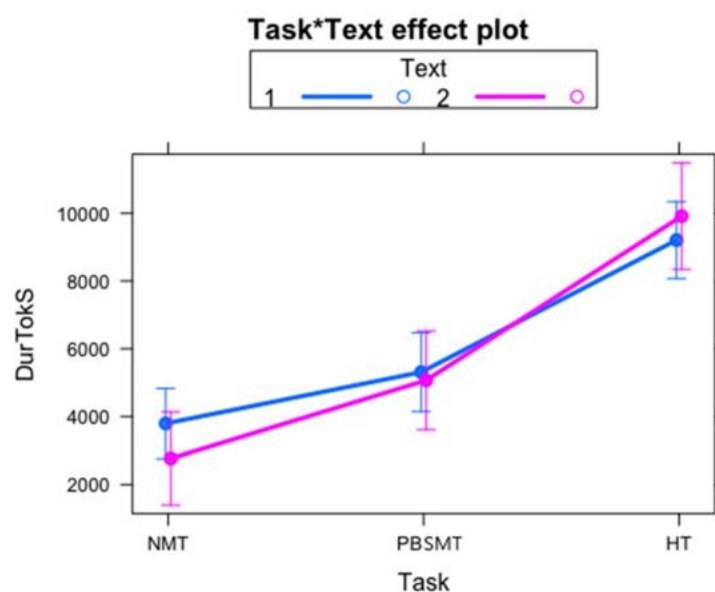


Fig. 3 Effect plot of interaction effect between task (HT, post-editing NMT and post-editing PBSMT) and text (Text 1 and Text 2) on average duration per token in ms (DurTokS)

Appendix IV Post-editing and human translation productivity (Huang & Wang, 2022)

PERSPECTIVES 13

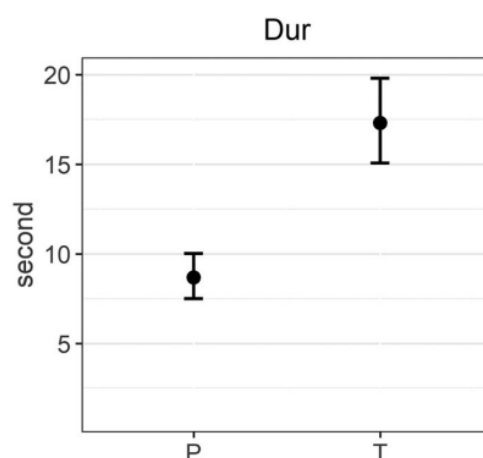
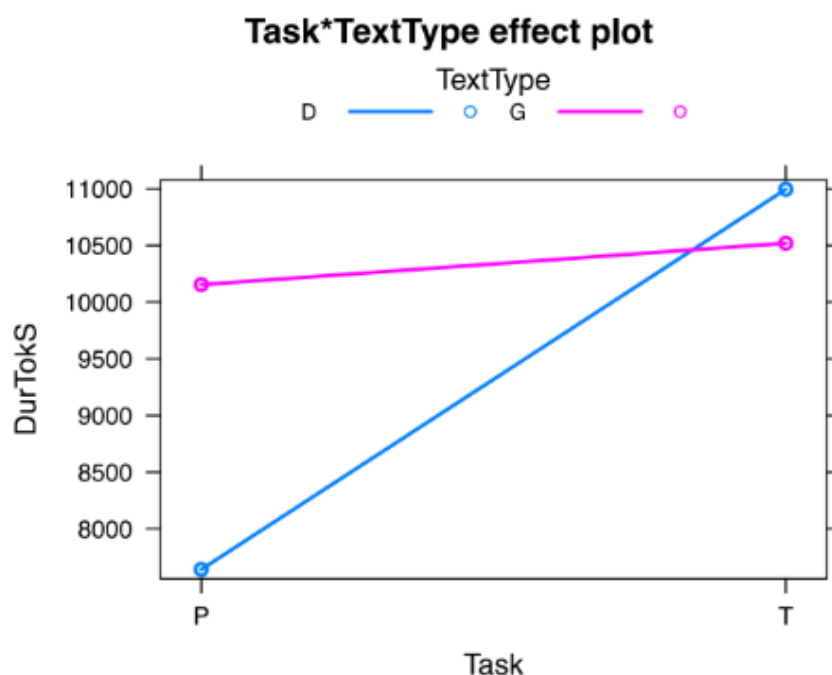


Figure 4. Effect plot of subtitling method on production duration (back-transformed to original scale).

Appendix V Post-editing and human translation productivity (Jia, et al., 2019b)



Appendix VI Post-editing and human translation productivity (Moorkens, et al., 2018)

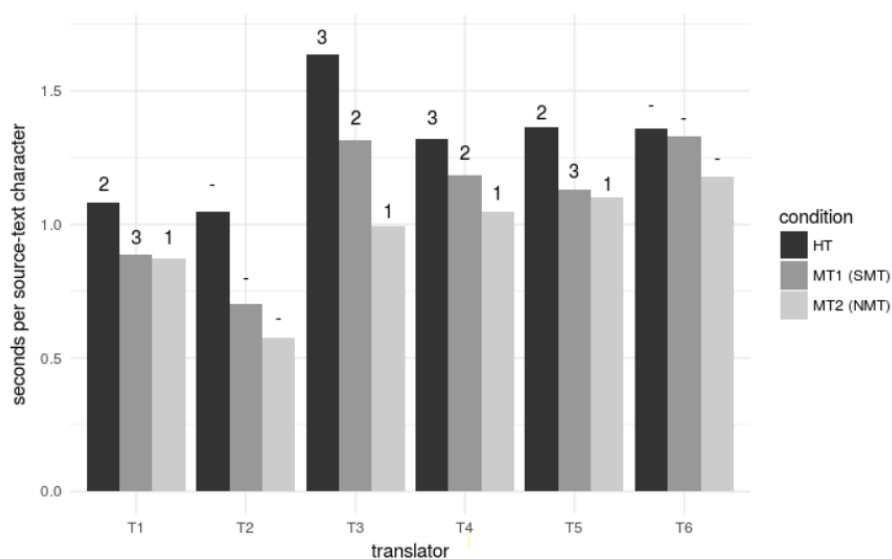


Figure 1. Measurements (displayed by bar height) and perceptions (displayed by rank) of temporal effort.

Appendix VII Productivity gains and no quality changes (Stasimioti & Sosoni, 2020)

Task	Mean	SD
Translation task	9.86	4.53
PE task	7.91	2.48

Table 10. Temporal effort per task: Mean and standard deviation values of the task duration

Task	Participants	Mean	SD
PE	Professionals with experience in PE	7.07	1.34
	Professionals without experience in PE	10.42	4.09

Table 11. Professionals with experience in PE vs professionals without experience in PE: Mean and standard deviation values of the PE task duration

Appendix VIII Productivity gains and no quality changes (Ahmad, et al., 2018)

Translator	1	2	3	4	5
Lang. Pair	Eng-Hin	Eng-Hin	Eng-Hin	Hin-Urd	Eng-Hin
Level	Noice	Beginner	Medium	Expert	Expert
Manual	500	700	1,000	1,200	1,200
Week-1	642	941	1,251	1,917	1,220
Week-2	770	863	1,158	2,456	2,040
Week-3	588	775	1,221	2,555	1,605
Week-4	555	949	1,362	2,658	1,746
Week-5	463	1,098	1,408	2,653	2,054
Week-6	648	995	1,449	2,626	2,153
Week-7	807	990	1,468	2,706	2,193
Week-8	818	940	1,552	2,680	2,255
Week-9	736	964	1,521	2,799	2,518
Week-10	828	903	1,508	2,779	2,735
Week-11	842	853	1,439	2,876	2,874
Week-12	934	860	1,452	2,975	3,049
Week-13	934	902	1,538	3,041	3,155
Week-14	980	968	1,584	3,129	3,252
Week-15	963	1,111	1,623	-	3,208
Week-16	964	1,209	1,614	-	3,252
Week-17	981	1,212	1,624	-	3,254
Week-18	953	1,187	1,640	-	3,215
Week-19	942	1,237	1,659	-	3,172
Week-20	930		1,633	-	3,120
Maximum	981	1,237	1,659	3,129	3,254
Gain (X)	1.96	1.77	1.66	2.61	2.71

Table 1: Shows Translators Productivity

Appendix IX language codes²⁰

CODE	LANGUAGE
AR	Arabic
CA	Catalan
DA	Danish
DE	German
EL	Greek
EN	English
ES	Spanish

²⁰ These codes were created by adapting the ISO 639-1 'Codes for the Representation of Names of Languages'.

FI	Finnish
FR	French
HI	Hindi
IT	Italian
NL	Dutch
PL	Polish
PT	Portuguese
SL	Slovenian
SV	Swedish
TR	Turkish
UK	Ukrainian
UR	Urdu
VI	Vietnamese
ZH	Chinese

Appendix X Productivity gains and no quality losses (Cui, et al., 2023)

Table 1. STW of HT and PE for different texts.

First, there is significant difference between HT and PE in terms of STW. As shown in Table 1, for the advertising text, STW of HT ($M = 5.017$, $SD = 1.731$, $ME = 4.757$, 95% CI [4.403, 5.631]) is higher than that of PE ($M = 3.105$, $SD = 1.600$, $ME = 2.882$, 95% CI [2.538, 3.672]), and such difference is statistically significant ($z = 4.119$, $p < .001$). The effect size is large ($r = .717$). For the news text, STW of HT ($M = 4.259$, $SD = 1.878$, $ME = 3.778$, 95% CI [3.594, 4.925]) is higher than that of PE ($M = 2.410$, $SD = 1.171$, $ME = 2.406$, 95% CI [1.995, 2.826]), and such difference is statistically significant ($z = 4.512$, $p < .001$). The effect size is large ($r = .786$). For the legal text, STW of HT ($M = 5.054$, $SD = 1.467$, $ME = 4.976$, 95% CI [4.534, 5.574]) is higher than that of PE ($M = 2.777$, $SD = 1.326$, $ME = 2.308$, 95% CI [2.306, 3.247]), and such difference is statistically significant ($z = 4.976$, $p < .001$). The effect size is large ($r = .866$). For the poetic text, STW of HT ($M = 6.804$, $SD = 2.173$, $ME = 6.467$, 95% CI [6.034, 7.575]) is higher than that of PE ($M = 4.748$, $SD = 1.758$, $ME = 4.956$, CI [4.124, 5.371]), and such difference is statistically significant

Correlation of Subjective and Objective Measures. Having analyzed the three dimensions of effort, we further explored the correlation of the subjective and objective measures. First of all, we compared participants' AS of HT and PE.

There is significant difference between the participants' scores of HT and PE. As shown in Table 3, for the advertising text, the AS of HT ($M = 84.384$,

$SD = 4.410$, $ME = 84.667$, 95% CI [82.820, 85.947]) is lower than that of PE ($M = 89.051$, $SD = 1.272$, $ME = 90.333$, 95% CI [89.599, 90.502]), and such difference is statistically significant ($z = -4.147$, $p < .001$). The effect size is large ($r = .818$). For the news text, the AS of HT ($M = 84.283$, $SD = 3.964$, $ME = 85$, 95% CI [82.877, 85.689]) is lower than that of PE ($M = 88.202$, $SD = 2.673$, $ME = 88.333$, 95% CI [87.254, 89.150]), and such difference is statistically significant ($z = -4.700$, $p < .001$). The effect size is large ($r = .721$). For the legal text, the AS of HT ($M = 86.202$, $SD = 2.586$, $ME = 86.667$, 95% CI [85.285, 87.119]) is higher than that of PE ($M = 85.758$, $SD = 1.444$, $ME = 86$, 95% CI [85.246, 86.270]), and such difference is statistically insignificant ($z = 1.198$, $p = .231$). The effect size is moderate ($r = .208$). For the poetic text, the AS of HT ($M = 78.808$, $SD = 5.311$, $ME = 77.667$, 95% CI [76.924, 80.691]) is lower than that of PE ($M = 83.525$, $SD = 3.299$, $ME = 83.333$, 95% CI [82.356, 84.695]), and such difference is statistically significant ($z = -4.137$, $p < .001$). The effect size is large ($r = .720$). It can be seen that the participants' transla-

Appendix XI Productivity gains and quality losses (Zhang & Qian, 2023)

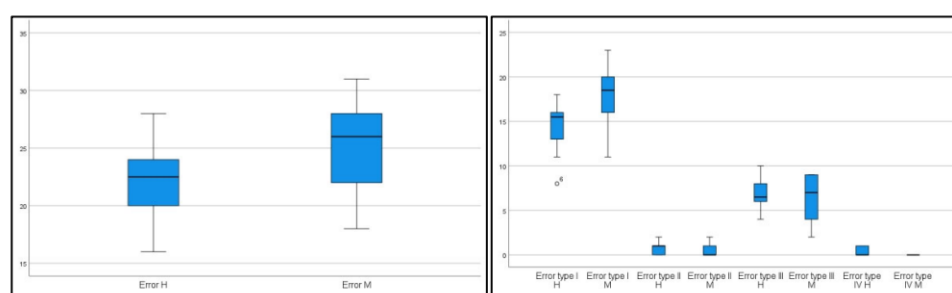
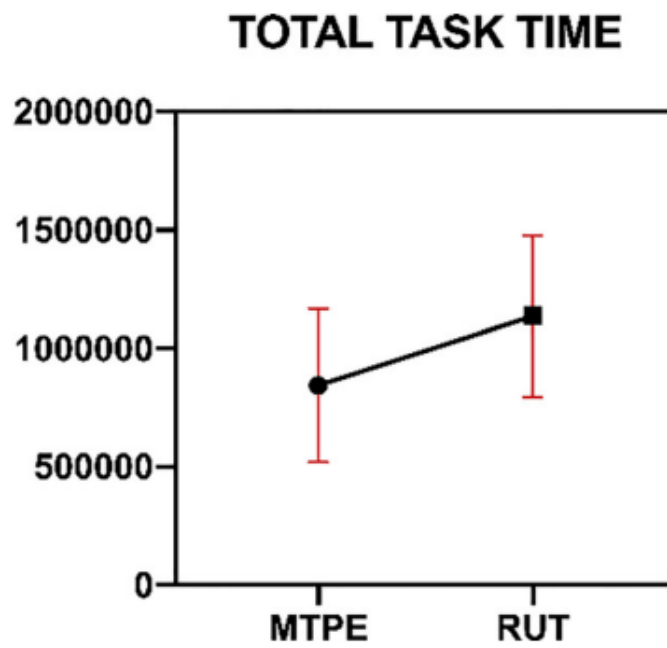


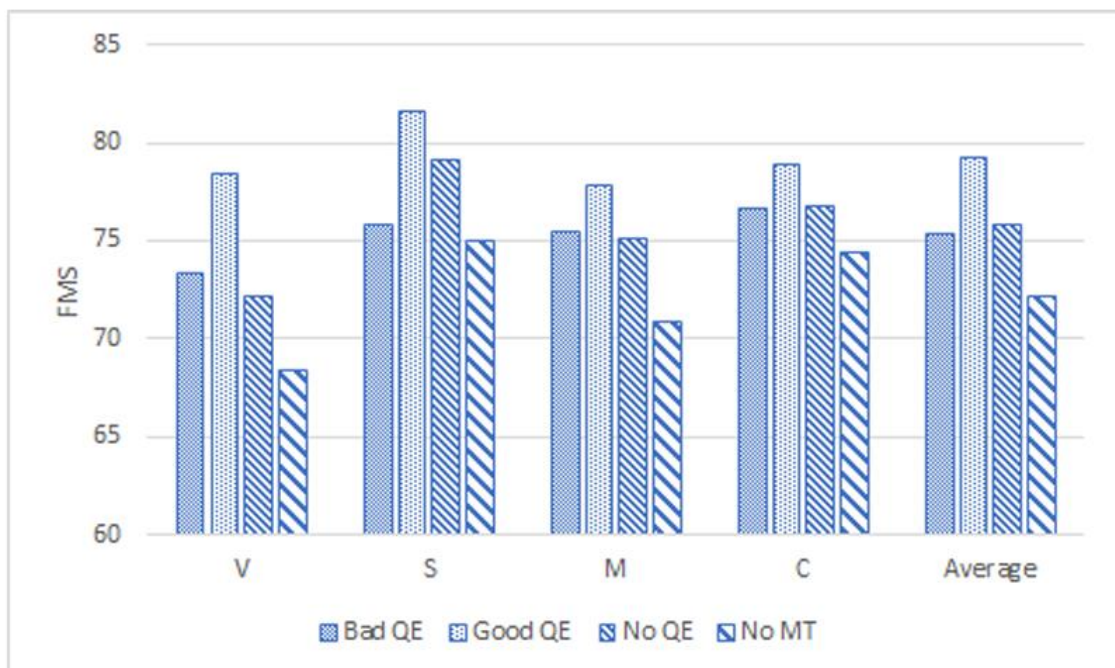
Figure 3 Boxplots of the error counts

The total error count in the from-scratch translations ($M = 22.30$, $SD = 3.368$) was lower than it was in the PEMT-assisted translations ($M = 24.80$, $SD = 4.341$), but not significantly so, with $t(9) = -1.90$, $p = .09$. Of the four types of errors, only type I accuracy errors in the from-scratch translations ($M = 14.4$, $SD = 3.169$) were significantly lower than they were in the PEMT-assisted translations ($M = 17.90$, $SD = 3.725$), with $t(9) = -2.71$, $p = .02$.

Appendix XII Productivity and quality gains (Wang, et al., 2021)



Appendix XIII Productivity and quality gains (Béchara, et al., 2021)



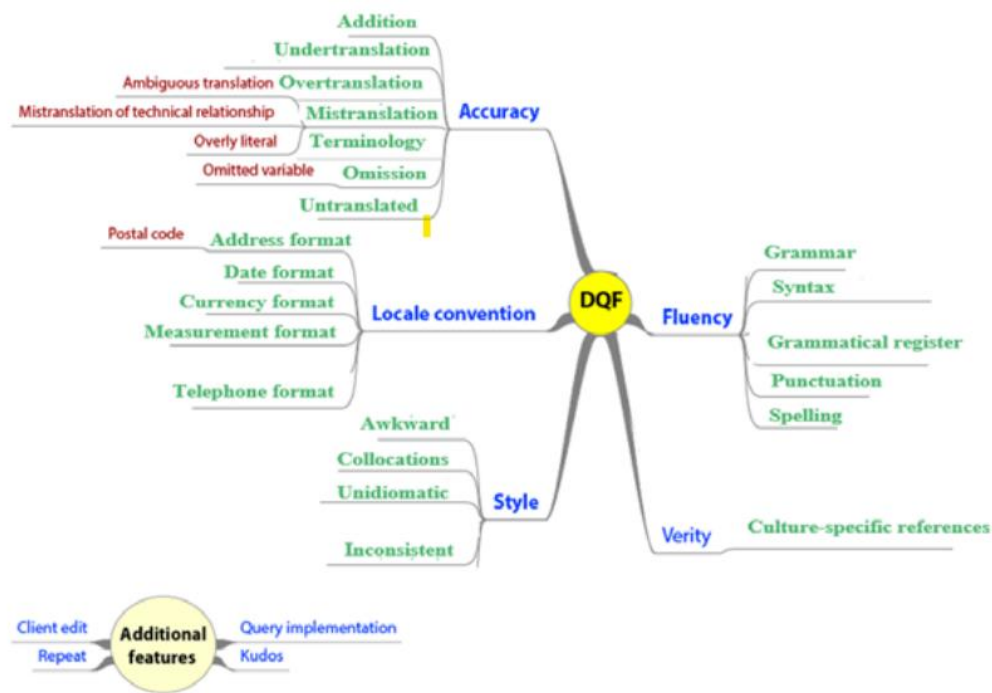
Appendix XIV Productivity and quality gains/losses²¹ (Kuzman, et al., 2019)

	Group A (person 1)– GNMT	Group A (person 2)– GNMT	Group B (person 1)– GNMT	Group B (person 2)– GNMT	Group C (person 1)– SPOOK	Group C (person 2)– SPOOK	Group D (person 1)– SPOOK	Group D (person 2)– SPOOK	Group E (person 1)–Novel	Group E (person 2)–Novel	Group F (person 1)–Novel	Group F (person 2)– Novel
Translation time (min)	9.5	12.6	6.0	12.3	17.2	8.4	12.8	14.0	11.7	17.6	10.7	12.1
PE time (min)	12.8	13.0	9.7	16.3	9.8	15.6	13.2	15.9	13.0	21.4	10.6	11.1
Difference between translation and PE time (min)	3.2	0.4	3.8	4.0	-7.4	7.1	0.4	1.9	1.3	3.8	-0.1	-1.0
The task that participants perceived to be more time-con- suming	translation	translation	PE	PE	translation	translation	PE	translation	translation	PE	PE	translation

Table 3. Measurement of the temporal dimension of post-editing effort

	GNMT	SPOOK	Novel
Average difference between translation and PE time (min)	2.9	0.5	1.0

Appendix XV Modified MQM-DQF (Sosoni, et al., 2022)



²¹ The authors, after being contacted on July 19th, 2024, confirmed, that the captions in the able are in a wrong order; therefore, the first row refers actually to PE time, and the second row to translation time.

Appendix XVI Abstract English

Daily, machine translation (MT) is used to produce great amounts of content (Lommel & Pielmeier, 2021), with post-editing (PE) serving as a key process in optimizing its output. However, the relationship between PEMT quality and productivity, as well as other influencing factors, remains unclear.

This thesis addresses these gaps by answering the following research questions: RQ1: What is the relationship between translators' productivity when post-editing machine translation and the translated text's quality? RQ2: Do PEMT productivity and quality differ according to text types, language pairs, types of participants and their PE experience, used NMT systems, and experiment settings? In order to answer these questions a text corpus consisting of thirty articles investigating PEMT productivity and quality in the case of NMT was analyzed. The data was organized in tables for RQ1 and in charts for RQ2.

Some of the main results of this exploration indicate that a) PEMT usually shows the highest productivity gains when the claimed unchanged quality is not formally measured; b) if the quality is measured, the productivity gains are much lower. For RQ2, findings indicate: a) some language pairs show a tendency to more gains/losses; b) professional translators and experienced participants exhibit the best scores; c) PE influences the most translations of students; d) commercial NMT systems yield the most consistent outcomes, while fine-tuning does not always guarantee stable results; e) PE general texts brings the most consistent outcomes; f) regular translation environment does not guarantee the best scores.

This thesis highlights that both PEMT productivity and quality are influenced by various factors, contributing to a deeper understanding of the ambiguity of studies conducted on this topic.

Appendix XVII Abstract German

Die maschinelle Übersetzung (MT) wird täglich genutzt, um große Mengen an Inhalten zu produzieren (Lommel & Pielmeier, 2021), wobei das Post-Editing (PE) als Schlüsselprozess zur Optimierung des Outputs dient. Allerdings ist der Zusammenhang zwischen PEMT-Qualität und Produktivität sowie anderen Einflussfaktoren noch unklar.

Diese Arbeit verfolgt das Ziel, diese Aspekte zu untersuchen, indem auf folgende Fragen eingegangen wird: RQ1: Welcher Zusammenhang besteht zwischen der PE-Produktivität bei der maschinellen Übersetzung und der Qualität des übersetzten Textes? RQ2: Wie hängen die Produktivität und die Qualität der PEMT mit der Textsorte, den Sprachpaaren, der Ausbildung der Teilnehmer:innen und ihrer Erfahrung mit PE, den verwendeten NMT-Systemen und der Versuchsanordnung zusammen? Um diese Fragen zu beantworten, wurde ein Textkorpus, der aus dreißig Artikeln besteht, analysiert. Die Artikel untersuchen die PEMT-Produktivität und -Qualität im Falle von NMT. Die Ergebnisse wurden für RQ1 in Tabellen und für RQ2 in Diagrammen dargestellt. Einige der wichtigsten Ergebnisse dieser Untersuchung zeigen, dass a) PEMT in der Regel die höchsten Produktivitätsgewinne aufweist, wenn die geforderte unveränderte Qualität nicht formell gemessen wird; b) wenn die Qualität gemessen wird, sind die Produktivitätsgewinne viel geringer. In Bezug auf Frage 2 zeigen die Ergebnisse, dass: a) einige Sprachpaare eine Tendenz zu mehr Gewinnen/Verlusten aufweisen; b) professionelle Übersetzer:innen und erfahrene Teilnehmer:innen die besten Ergebnisse erzielen; c) PE die meisten Übersetzungen von Student:innen beeinflusst; d) kommerzielle NMT-Systeme die konsistentesten Ergebnisse liefern, während Fein-Tuning nicht immer stabile Ergebnisse garantiert; e) PE allgemeiner Texte die konsistentesten Ergebnisse erzielt; f) reguläre Übersetzungsumgebung nicht die besten Resultate garantiert.

Die vorliegende Arbeit zeigt auf, dass sowohl die Produktivität als auch die Qualität von PEMT von verschiedenen Faktoren beeinflusst werden, was zu einem tieferen Verständnis der Mehrdeutigkeit der zu diesem Thema durchgeführten Studien beiträgt.