Hacettepe University Graduate School of Social Sciences

Department of Translation and Interpreting

# TEMPORAL AND TECHNICAL EFFORT IN POST-EDITING COMPARED TO EDITING AND TRANSLATION FROM SCRATCH

Volkan DEDE

Master’s Thesis

Ankara, 2022

# TEMPORAL AND TECHNICAL EFFORT IN POST-EDITING COMPARED TO EDITING AND TRANSLATION FROM SCRATCH

Volkan DEDE

Hacettepe University Graduate School of Social Sciences
Department of Translation and Interpreting

Master's Thesis

Ankara, 2022

# KABUL VE ONAY

Volkan Dede tarafından hazırlanan “Temporal and Technical Effort in Post-editing Compared to Editing and Translation from Scratch” başlıklı bu çalışma, 26.05.2022 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Prof. Dr. Sultan Çiğdem SAĞIN ŞİMŞEK (Başkan)

Doç. Dr. Elena ANTONOVA ÜNLÜ (Danışman)

Dr. Öğr. Üyesi Alper KUMCU (Üye)

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof.Dr. Uğur ÖMÜRGÖNÜLŞEN
Enstitü Müdürü

# YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinleri yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- o Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. [(1)]
- o Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ….. ay ertelenmiştir. [(2)]
- o Tezimle ilgili gizlilik kararı verilmiştir. [(3)]

16/06/2022

**Volkan DEDE**

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

(1) *Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

(2) *Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

(3) *Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
*Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.*

** Tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

# ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, **Doç. Dr. Elena ANTONOVA ÜNLÜ** danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

**Volkan DEDE**

# ACKNOWLEDGEMENTS

> "People will never be replaced by machines. In the end, life and business are about human connections. And computers are about trying to m-rder you in a lake. And to me, the choice is easy." – Michael Scott, *The Office*

I have been obsessed with machine translation since I became aware of it. And don't fear, it will not replace us translators any time soon – and we as linguists should make every effort to understand and adopt our newest aid.

This thesis took the longest time to write, thanks in part to the pandemic and my endless procrastination. Yet, in the end, it's finally here. I've never thought that I would be able to finish it in the midst of the most severe depressive episode I've had over the years, and I want to note it here so that years later, I can check this page and realize what I'm capable of – with the valuable help of my professors and my friends whom I'd like to thank.

First and foremost, I'd like to express my fondest gratitude to Assoc. Prof. Dr. Elena Antonova Ünlü because she has never given up on me even though I'm sure none of her students had to take 4 years to complete a master's thesis. She has always provided me with valuable insight and delicious coffee and pushed me to be better.

I also want to thank Asst. Prof. Dr. Sinem Sancaktaroğlu Bozkurt for her support during this period and for frequently reminding me through direct messages that I had a thesis to write instead of watching TV shows. I would also like to thank Assoc. Prof. Dr. Elif Ersözlü and Asst. Prof. Dr. Alper Kumcu for their support particularly during the inception phase of this thesis. I'm grateful to all other faculty members at my department because they've contributed to this thesis in ways they do not know.

I want to thank each and every single one of the students who took part in my experiment because without them, I'd have no data.

I want to thank my friends, especially those who were going through a similar dreadful process of finishing their theses (and did so before me), and for all their support and encouragement: Başak Şen, Gözde Yeşil, Sena Ekinci, Didar Sevde Ülker, Gülce Naz Şemi, Doğukan Duman, Esra Misem Sönmez, Melisa Doğan, and Nagehan Sayıcı.

I want to thank my cats, Kiki and Miki, who never did anything to break my new laptop after I had lost the last one to “fluid contact.”

# ÖZET

DEDE, Volkan. *Post-editing Eyleminde Zamansal ve Teknik Eforun Düzeltme ve Sıfırdan Çeviri ile Karşılaştırılması,* Yüksek Lisans Tezi, Ankara, 2022.

Bu tezin amacı, makine çevirisi post-editing eyleminde çevirmenlerin makine çevirisi çıktılarını düzenlemek için harcadığı zamansal ve teknik eforu düzeltme ve sıfırdan çeviri eylemleriyle karşılaştırmaktır. Araştırma aynı zamanda tez kapsamında özel olarak eğitilen bir istatistiksel makine çevirisi motoru ile ücretsiz bir nöral makine çevirisi motorunu kıyaslamaktadır. Tezin amaçları doğrultusunda, yükseköğretim öğrencilerinden oluşan bir örneklem; İngilizce-Türkçe dil çiftinde makine çevirisi çıktılarını düzenlemeleri, insan çevirisini düzeltmeleri veya sıfırdan çeviri yapmaları gerektiği bir deneye tabi tutulmuştur. Deney, yaygın bir bilgisayar destekli çeviri aracı üzerinde haber metinleriyle gerçekleştirilmiştir. Katılımcıların cümleleri düzenlerken harcadıkları zaman ve yaptıkları düzenleme miktarı nicel olarak ölçülmüştür. Araştırma sonucunda, özel eğitilen istatistiksel makine çevirisi motoru ile nöral makine çevirisi motoru arasında anlamlı bir farklılık bulunmamıştır. Katılımcıların sıfırdan çeviri ve insan çevirisinin düzenlenmesi sırasında, makine çevirisine göre daha fazla teknik ve zamansal efor harcadıkları bulunmuştur. Bu tez, çeviri sektöründe nispeten yeni bir hizmet olan post-editing'in müşteriye, projeye ve çevirmene yararı değerlendirilirken ilgili paydaşlara bir rehber olmayı ve bu dil çiftinde ihtiyaç duyulan benzer çalışmaları teşvik etmeyi amaçlamaktadır.

**Anahtar Sözcükler**

makine çevirisi, postediting, düzeltme, zamansal efor, teknik efor

# ABSTRACT

DEDE, Volkan. *Temporal and Technical Effort in Post-editing Compared to Editing and Translation from Scratch*, Master's Thesis, Ankara, 2022.

The aim of this thesis is to compare the temporal and technical effort spent by translators to post-edit machine translation outputs with editing and translation from scratch. The research also compares a statistical machine translation engine specially trained for the experiment with a public neural machine translation engine. For the purposes of the thesis, a sample of higher education students took part in an experiment in which they had to post-edit machine translation output, edit human translation, or translate from scratch from English to Turkish. The experiment was conducted with news texts on a common computer-assisted translation tool. The amount of time participants spent editing sentences and the amount of editing they did were quantitatively measured. The results showed that there was no significant difference between the specially trained statistical machine translation engine and the neural machine translation engine. It was found that the participants spent more technical and temporal effort when translating from scratch and editing human translation than post-editing machine translation. This thesis aims to serve as a guide for stakeholders in evaluating the benefits of post-editing, a relatively new service in the translation industry, for the client, the project and the translator, and to encourage much-needed similar studies in this language pair.

**Keywords**

machine translation, postediting, editing, temporal effort, technical effort

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION

## i. BACKGROUND OF THE STUDY

Globalisation has transformed translation into a necessity for companies in every industry around the world willing to cross borders. For instance, the streaming boom in the entertainment industry or the decentralised procedures allowing more global clinical trials to be conducted has made speed a critical part of the professional translation process for various industries, and translators working in different language pairs have begun to face an ever-increasing volume of projects.

Translation technologies have been a matter of interest since the Cold War era, when the first attempts at fully automatic translation systems were made because of the scarcity of human translators working in the Russian-English language pair or lack of trust in them due to potential espionage. In the present day, however, the primary reason for the desire to integrate technology into the translation process is to increase the productivity of translators, speed up the process, and meet the demand.

Corresponding to the necessity of translation in different industries, machine translation has come into play to aid the human translator in the translation process. Machine translation is expected to increase productivity and quality and is quickly becoming a common service provided by language service providers worldwide.

Although the search for an automatic translation system has been continuing for years, the editing of machine translation outputs by human translators or the process thereof is a relatively new concept waiting to be elucidated.

As it is more commonly called, postediting is a complex process situated in between traditional editing and translation from scratch. Both internal and external factors play a role in the efficiency of postediting, such as the quality of the machine translation engine, the technological competence of the human translator, or the language pair used.

Such various factors at play require that specific investigations be carried out for different language pairs, engines, and text types.

## ii. PURPOSE AND SCOPE OF THE STUDY

This thesis relates to the effort exerted in machine translation postediting, which involves human-translators editing raw outputs generated by machine translation systems.

Several indicators of machine translation postediting effort are used to elucidate the process. To this end, translation students at postgraduate levels were enrolled to take part in an experiment. The participants with varying degrees of professional experience in the translation industry were asked to work on four types of segments: human translation (translation memory matches), raw outputs from a statistical machine translation system, raw outputs from a neural machine translation system, and empty segments to be translated from scratch. Pre- and post-test questionnaires were applied to obtain information about the participants and their perceived effort of the task. Processing time on each segment was recorded, and edit distance was calculated post-experiment.

This thesis is motivated by the rigorous body of research into the translation process, which Krings (2001) has comprehensively compiled in a review to then extrapolate their findings to the study of postediting processes. The distinction between the two processes is said not to be crystal clear, as postediting triggers the same reading processing involved in reading any other text type (Krings, 2001); on the other hand, the raw target text in postediting is characterized by linguistic and semantic deficiency (Krings, 2001) in that the translation is incomplete and requires additional treatment from the human-translator in order to be accurate and adequate. Therefore, the question remains whether the process of postediting significantly differs from that of translation or editing.

Three main types of effort are investigated in traditional postediting process research, which are based on Krings's (2001) classification: temporal, technical, and cognitive effort. While the first two is directly quantifiable, the direct measurement for the last one is assumed to be impossible (Krings, 2001). Therefore, cognitive effort can only be measured through indirect approaches such as Think-Aloud Protocols (TAPs), subjective effort scales, keyboard logging, or with a combination of the first two types of effort (Koponen, 2012; Lacruz et al., 2012; Popović et al., 2014; Vieira, 2016 inter alia).

Investigations into the postediting process is essential in order to find answers to several questions, from both academic and industrial points of view, including the necessary skills and training for a good postediting performance and fair payment practices for postediting services in the industry. As such, this thesis attempts to provide the much-needed answers to these questions in the context of Turkish translation literature.

### iii. RESEARCH QUESTIONS

As stated above, this thesis aims to offer insight into the process of postediting in the English-Turkish language pair. Specifically, this thesis seeks to answer the following questions:

1. Do temporal and technical effort in postediting tasks show significant differences as compared to editing and translation tasks?

2. Do temporal and technical effort in postediting tasks differ between a statistical machine translation engine and a neural machine translation system?

3. Does the subjective effort of post-editors relate to their actual measured effort?

The aforementioned research questions were investigated with an extensive experiment in a controlled environment designed considering what has been achieved so far in the studies summarized in the following chapters.

### iv. SIGNIFICANCE OF THE STUDY

Postediting is a relatively new concept in the translation industry and an even newer service when the local language service providers in Turkey are considered. In addition, postediting has rarely been studied in an academic setting for the English-Turkish or Turkish-English language pairs. Therefore, this thesis aims to fill the gap in the Turkish literature about the process of postediting with an additional focus on providing the stakeholders in the industry with much-needed answers to their questions about postediting effort, which could help in decisions related to implementation of machine translation, pricing of postediting, and management of postediting projects.

## v. LIMITATIONS

Process-oriented studies in the translation field face many problems to overcome. Subjective factors such as the speed of translation, attitude towards or competence in translation technologies complicate the efforts to elucidate the process. Text type, language pair, and the choice of computer-assisted translation tool have an additional impact on the overall process of translation and postediting. All possible efforts were made to reduce the impact of such factors in the present study, such as the selection of a common translation tool and enrolment of a similar population of translators.

This study is limited to one language pair and uses a public corpus comprising news texts. While the choice of using news texts was practical for the purposes of the experiment, it is a relatively uncommon text for professional translators to work on.

Since the conduct of the experiment, statistical machine translation has somewhat become obsolete, and newer, more successful neural machine translation engines have been introduced, while this thesis used a neural engine that has remained common.

There are three main types of effort investigated in the postediting process, and this thesis only deals with temporal and technical effort. Direct quantification of the third type, cognitive effort, might involve the use of measurement methods interfering with the translator's working environment, such as eye-tracking, and the researcher had no access to such tools and wanted the environment to resemble a typical atmosphere for a translator.

English proficiency was not formally tested due to the assumption that the students taking part in the experiment had a good command of the language since they were enrolled in the English translation programme.

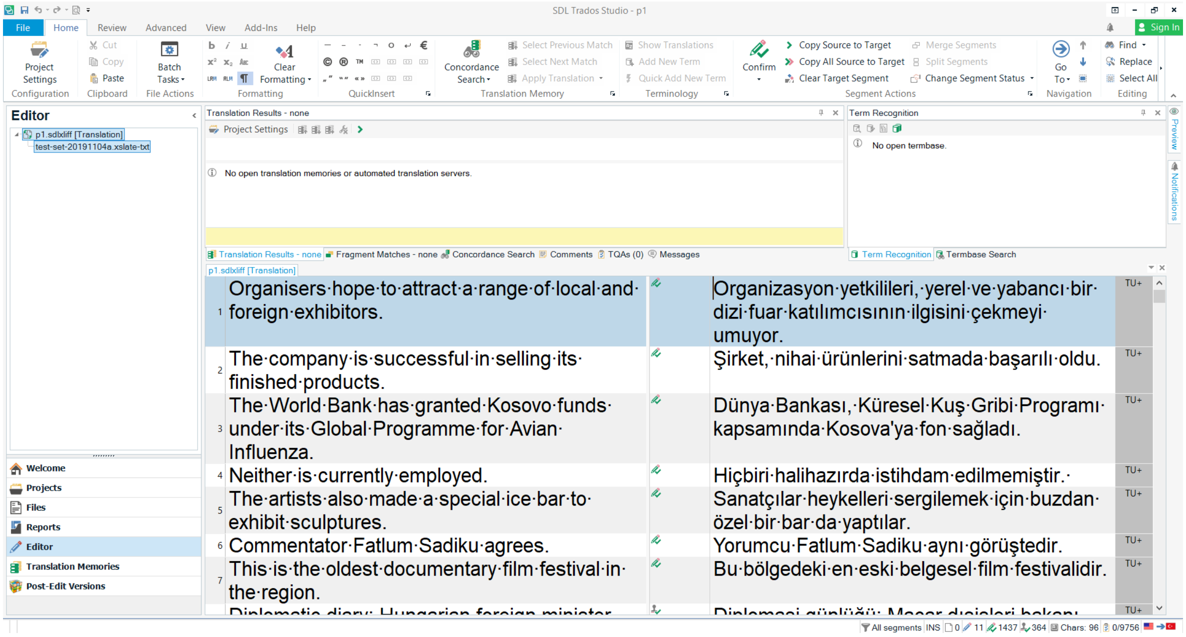
# CHAPTER 1

# LITERATURE REVIEW

## 1.1. COMPUTER-ASSISTED TRANSLATION

Technological advances have revolutionized the way translators work  (Folaron, 2010). In as early as 1980, Martin Kay (as cited in Schwartz, 2018) predicted how computers could transform the translation profession, which led him to suggest a cooperation, rather than competition, between humans and machines. The revolutionary developments in computer technologies introduced many tools to the workstation of the translator, collected under the term "computer-assisted translation (CAT) tools."

*Figure 1. Example of a CAT tool's user interface (SDL Trados 2017)*



No open translation memories or automated translation servers.

No open termbase.

| | | | |
|---|---|---|---|
| 1 | Organisers·hope·to·attract·a·range·of·local·and·foreign·exhibitors. | Organizasyon·yetkilileri,·yerel·ve·yabancı·bir·dizi·fuar·katılımcısının·ilgisini·çekmeyi·umuyor. | TU+ |
| 2 | The·company·is·successful·in·selling·its·finished·products. | Şirket,·nihai·ürünlerini·satmada·başarılı·oldu. | TU+ |
| 3 | The·World·Bank·has·granted·Kosovo·funds·under·its·Global·Programme·for·Avian·Influenza. | Dünya·Bankası,·Küresel·Kuş·Gribi·Programı·kapsamında·Kosova'ya·fon·sağladı. | TU+ |
| 4 | Neither·is·currently·employed. | Hiçbiri·halihazırda·istihdam·edilmemiştir.· | TU+ |
| 5 | The·artists·also·made·a·special·ice·bar·to·exhibit·sculptures. | Sanatçılar·heykelleri·sergilemek·için·buzdan·özel·bir·bar·da·yaptılar. | TU+ |
| 6 | Commentator·Fatlum·Sadiku·agrees. | Yorumcu·Fatlum·Sadiku·aynı·görüştedir. | TU+ |
| 7 | This·is·the·oldest·documentary·film·festival·in·the·region. | Bu·bölgedeki·en·eski·belgesel·film·festivalidir. | TU+ |
| | Diplomatic·diary:·Hungarian·foreign·minister | Diplomasi·günlüğü:·Macar·dışişleri·bakanı | TU+ |

Starting with simple word processors, computer-assisted translation tools evolved to something more sophisticated towards the end of the 20th century. With the first commercial computer-assisted translation tool (Translation Support System) developed in the 1980s, translation technologies witnessed a fast rise (Sin-wai, 2017) in development and adoption. The lack of success obtained in efforts to automate translation (discussed in the next section) led scholars like Bar-Hillel (as cited in Poibeau, 2017) to suggest shifting the technological works towards computer-aided translation, rather than fully-automatic translation. Tools developed as add-ons to word processors (e.g. in the case of Trados [Sin-wai, 2017]) later evolved to stand-alone tools, such as Transit and Translator's Workbench (Sin-wai, 2017). As of 2020, there are more than 30 computer-assisted translation tools according to an online database (CAT Tools | Software Comparison Tool, n.d.).

Despite the abundance of translation tools available on the market, all share some common features:

1. Segmentation, where the source text (ST) is divided into segments so that the translator can focus on each translatable item in ST separately;
2. Translation memory (TM), which stores the translator's work to be used later for similar projects.
3. Glossary or termbase, where the client- or project-specific terminology can be entered, eliminating the need to check dictionaries during the translation process.
4. Machine translation (MT), which automatically translates the source text and eliminates the need for manual translation. The translator/editor only has to perform postediting. computer-assisted translation tools do not usually come with their own machine translation system but use external resources.
5. Quality assurance (QA), which allows the translator to check technical errors with a dedicated functionality instead of performing them manually. Quality assurance checks can highlight errors related to spelling, punctuation, formatting, etc. with a possibility of customization for language-specific errors. Quality assurance can be integrated with the computer-assisted translation tool or can also be a standalone tool.

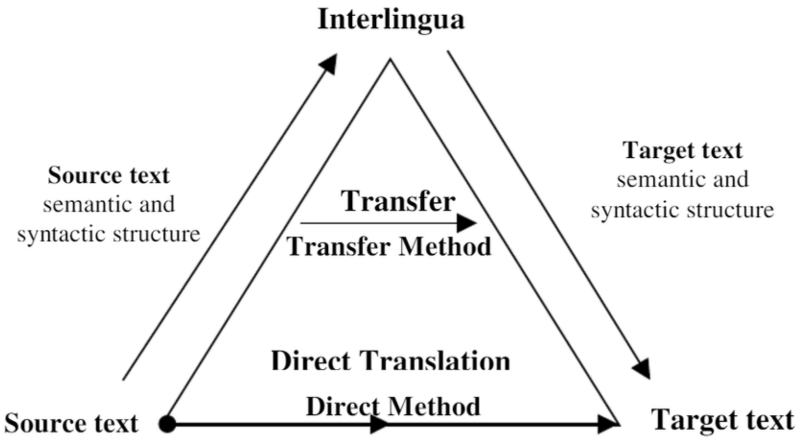
The listed are only five of the functionalities offered by different software. Depending on the brand, additional features may include more options for automation, cloud-based solutions, an increased amount of file type support, among others.

### 1.2. MACHINE TRANSLATION

Machine translation (MT) refers to the use of computers to produce translations of any given text rather than employing human-translators. Machine translation is based on cryptology and universal language theories, therefore its roots can be found in the works of 7$^{th}$-century Arabic cryptologist scholars (DuPont, 2018) or 17$^{th}$-century theorists of a universal language (W. J. Hutchins, 1986). When it became a reality, however, was during the Cold War era, when the parties, namely the United States of America and the Soviet Union at the time invested in technologies to achieve mechanical translation between English and Russian. This effort was aimed towards "Fully-Automatic High-Quality Translation" (FAHQT; Hutchins, 1986). Two years after M.I.T's first conference on "Mechanical Translation" in 1952 (O'Brien, 2012), the first-ever machine translation system developed jointly by Georgetown University and International Business Machines (IBM) was demonstrated. The system could translate 60 Russian sentences into English (W. J. Hutchins, 1986) with a database of 250 words and 6 grammar rules (J. Hutchins, 1999). The latter feature of these pioneer systems led to their labelling as "rule-based systems."

Rule-based systems, or rule-based machine translation (RBMT), consisted of three main approaches: direct, interlingual, and transfer systems.

*Figure 2. Vauquois triangle (obtained from* Chemvura, 2017*)*

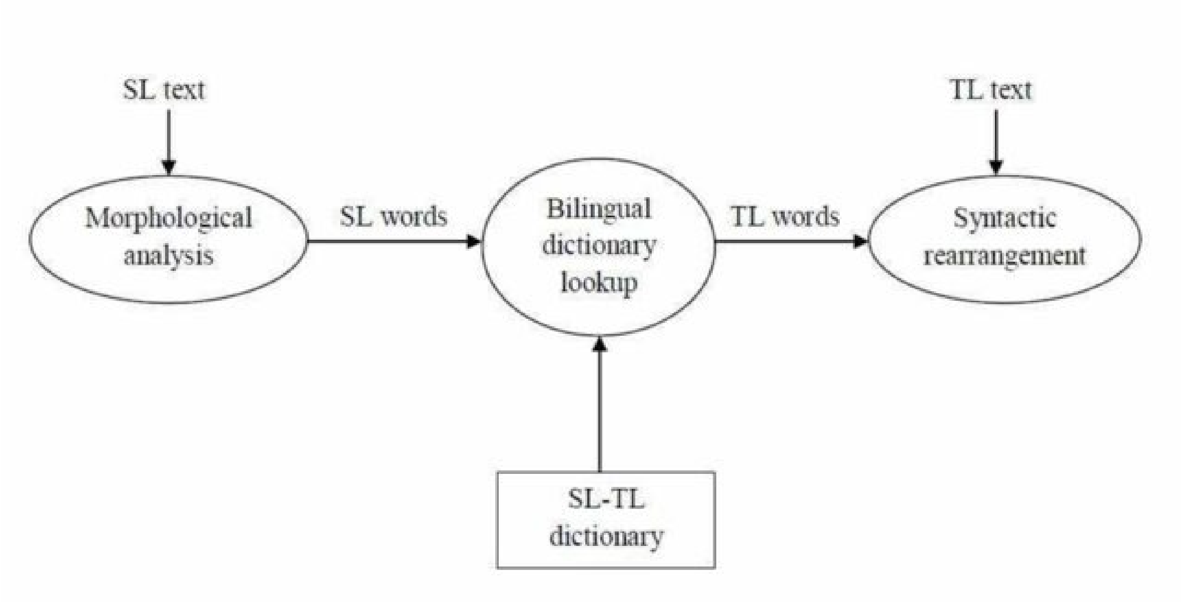


The triangle in Figure 2 summarizes the different levels of processing involved in varying approaches to machine translation well. As one goes up the triangle, the analysis becomes deeper; that is, the semantic and pragmatic functionalities of the respective system increases, while the bottom level is characterized by superficial comprehension. For instance, the "direct" approach, at the shallow end of the triangle, had two main components: a dictionary for source and target languages and a set of grammar rules. As Hutchins (1986, p. 54) puts it:

> "The basic assumption [in direct approach] is that the vocabulary and syntax of SL texts need not be analysed any more than strictly necessary for the resolution of ambiguities, the correct identification of appropriate TL expressions and the specification of TL word order."

Therefore, the source text would undergo minimal analysis, to the extent that it was required for a comprehensible target text. An algorithm for a direct translation system could only apply to a given language pair and in a single direction. This meant that a new system had to be developed for each language pair and each direction in that pair.

*Figure 3. The workflow of a direct machine translation approach (adapted from Sankaravelayuthan & Vasuki, 2013)*



Interlingual systems, on the other hand, depended on abstract representations, and being partly independent of language, the same algorithm could apply to several languages. While direct systems used a single source language-target language dictionary, interlingual systems had separate dictionaries and grammar rules for each. Compared to its predecessor, the interlingua approach would have the machine translate the text, not into the target language, but an interlingua, which would only then be transferred to the target language. Similarly, the last approach in the rule-based machine translation family, the transfer systems, built upon the interlingual approach in that they would also use representations, but the representations would be separate for the source and target language. There was also no claim of universality in transfer systems, as the higher amount of steps in the machine translation process required rules that could not be jointly applied to several languages at once (Hutchins, 1986).

Rule-based machine translation was succeeded by data-driven systems, more generally named as statistical machine translation (SMT) systems. These corpus-based approaches depended on a large set of aligned, parallel bilingual texts. Statistical machine translation is trained on such datasets in order to "learn" the language. Statistical machine translation eliminated the need for manual insertion of linguistic rules, as the machine could extract them from the dataset. Although statistical machine translation systems, much like rule-based machine translation, has different sub-approaches, the idea is simple:

> "The essence of the method is first to align phrases, word groups and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language." (Hutchins, 1999, p. 17)

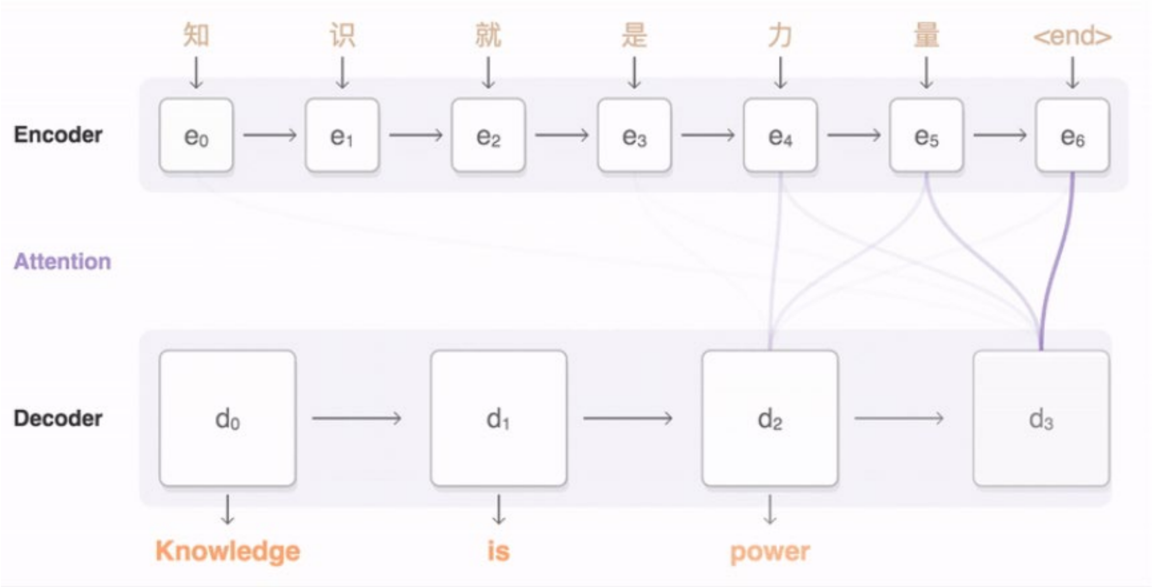
Simply put, if one were to compare the logic of statistical machine translation to rule-based machine translation, it could be said that statistical machine translation would generate the dictionaries and grammar rules (manually entered in rule-based machine translation approaches) with its training data. Another major difference of statistical machine translation from rule-based machine translation is that rule-based machine translation did not involve any corpus data for training. The level of data alignment (word or phrase) in statistical machine translation characterized the subcategories of statistical machine translation.

Phrase-based (or example-based) machine translation models (PBMT), for instance, would take pairs of phrases as atomic units (instead of single words) and were the common models employed by industry leaders (incl. Google Translate) until recently. The idea behind phrase-based models could be said to originate from Nagao's (as cited in Anastasiou, 2008) statements, where he suggested a similar system based on his claim that human translators worked by dividing the source sentence into fragments (i.e. phrases or words). Phrase-based machine translation’s main component to produce translations was "similarity scores" between fragments, while the original statistical machine translation systems worked on probability measurements. Later, the systems mentioned above would be combined to create hybrid systems (rule-based machine translation + statistical machine translation), where one could compensate for the deficiencies of the other approach, i.e. rule-based machine translation could be used to improve the grammatical aspect of statistical machine translation while statistical machine translation itself compensated for the semantic deficiencies of rule-based machine translation.

The latest development in corpus-driven machine translation technologies is neural machine translation (NMT). Although neural machine translation can be cited as a different approach to machine translation than the two described above, the fact that it also requires a bilingual corpus makes it reasonable for it to be listed under "corpus-driven" models. Neural machine translation has developed from efforts on artificial intelligence (AI) and deep/machine learning (ML). What distinguishes neural machine translation from earlier systems is that it can produce more fluent and accurate outputs than its antecedents. Neural machine translation has often been cited to produce "human-like" translations compared to its counterparts (Lilly, 2016).

The reason for the "humanness" of outputs generated by neural machine translation systems is the "neural" architecture, often claimed to mimic the human brain (Thames, 2019). Unlike rule-based machine translation or statistical machine translation, neural machine translation does not work from fragments or units but deals with the whole source sentence as one unit, or at least, this was the case in the earlier approaches to neural machine translation. Nowadays, most neural machine translation systems possess proprietary models, the common one being the "attention" model where neural machine translation returns to the fragmentation approach since it was later realized that neural machine translation would fail to demonstrate success when faced with longer inputs and often omit parts of sentences, resulting in incomplete translations. Another model is convolutional networks, helping the system dynamically process the sentence during the encoding process for acceleration.

*Figure 4. A representative schematic of a neural machine translation system (obtained from* Farooq, 2018*)*



The above figure demonstrates a simple schematic of how neural machine translation works. In this case, the Chinese characters (or words in Western languages) are encoded by the respective component into vectors, which are then decoded by the second component of the system. These are later transformed into concrete linguistic representations, rather than abstract vectors, and the final translation is produced.

Since the first demonstration of neural machine translation systems in 2016 (with Google's announcement; see Le & Schuster, 2016), interest in machine translation systems has skyrocketed with the amount of research increasing by 115% from 2017 to 2018 (Diño, 2018). Indeed, it is claimed that the last few years witnessed more advances in machine translation technologies than what had been achieved in the last ten years (Turovsky, 2016). The interest in machine translation-related services mirrored this trend, with the industry shifting towards machine translation postediting rather than human translation (Lommel, 2016). The demand for machine translation postediting services shows a steady gain according to a recent market research by CSA (Lommel, 2016). From an industrial viewpoint, machine translation postediting also provides "nearly 80% faster time to market at almost 80% less cost" (Milengo GmbH, 2019, para. 10). In an industry characterized by urgency, it can be said that "MT is here to stay" (CSA Research, n.d.).

## 1.3. POST-EDITING

### 1.3.1. Pre- and Postediting

Producing quality and often publishable outputs with machine translation is challenging, and from the earliest stages of machine translation development, human intervention was necessary both before and after the machine translation process. These acts were therefore termed "pre-editing" and "postediting," the latter of which concerns this thesis.

The prefixes indicate when the editing is supposed to be carried out on the source and/or target text. Pre-editing relates to revisions of the source text in order to make the text more suitable for computer processing. The extent of pre-editing depends on what is expected from the outcome. It can simply include dividing sentences, removing any ambiguities, while its most extreme form involves the use of "controlled language" (CL) rules (Gross, 1992). These rules dictate even which words are permitted in the source text for obtaining optimal efficiency from the machine. Pre-editing, in turn, can reduce the work that has to be done after the translation is complete, i.e. postediting. Indeed, pre-editing was found to have a significant effect on postediting by reducing the task time "almost by half" (Gerlach et al., 2013).

Postediting, on the other hand, concerns the target text and relates very closely to traditional editing as it involves the same basic steps with the exception of the author of the text, which is a machine. Postediting aims to bring the raw machine translation output closer to human-quality translation. The degree of postediting varies depending on several factors, including the purpose and type of the text, the intended audience, etc.

Postediting is carried out primarily at two levels: light or heavy/full postediting. Light postediting requires minimal editing on the target text, and its main purpose is usually "gisting," meaning that the text will not be disseminated outside a certain audience, therefore spending maximal effort is redundant. Light postediting can be applied, for example, to internal corporate documents that only a handful of people will read, or when quality is not the main concern of the translation project.

Heavy or full postediting, on the other hand, requires the utmost effort to make the text publishable, or similar or equal to a human quality translation. Full postediting involves a deeper examination of the raw output to identify all errors; semantic, linguistic, or grammatical. The editor then progresses to correct all identified deficiencies with the resulting text sounding as if it had been translated by a human. Indeed, Screen (2019) investigated the experiences of end-users when they were faced with a human translation and fully post-edited text and found that there was no negative effect on readability and perceptions of end-users induced by full postediting compared to human translation. The author concluded that, as there was no adverse quality or readability-related issues in fully post-edited texts, further integration of machine translation into professional workflows should be explored since it provides quality texts with a faster translation process.

The extent of postediting to be applied has drawn attention from both the industry and academia. Although postediting guidelines tend to be internal and specific to a given client or project type, several public instructions exist, specifying the right way to carry out postediting, the respective requirements, and essential considerations. A commonly cited one of those is the guidelines issued by TAUS (Translation Automation User Society), written by Massardo et al. (2016). TAUS uses the binary typology above with a different terming: good enough quality (equivalent to light postediting) vs. human translation quality (full postediting) (Massardo et al., 2016). For illustration purposes, the guidelines for full (human translation quality) postediting can be found below:

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

Ensure that formatting is correct.

(Massardo et al., 2016, p. 18)

The usability of these guidelines has been questioned by Flanagan and Christensen (2014), who set out to test the TAUS guidelines on translation students. They went on to suggest a new set of rules based on the comments of the participating trainee translators, who experienced frustration with some of the instructions given by TAUS. For instance, the instruction to "use as much of the raw machine translation output as possible" was placed on the top in the revised guidelines, as the study participants stated that they made some preferential changes (Flanagan & Christensen, 2014), as is expected for first-time post-editors (see Aranberri, 2017).

### 1.3.2. Process of Postediting

Postediting is regarded as a decision-making process, while from-scratch translation is a problem-solving task (Stefaniak, n.d.). In postediting, translators do not identify correct solutions but rather select from a pool of available solutions (Stefaniak, n.d.), which is what distinguishes postediting from translation or editing.

The potential differences in the process of postediting have led to a plethoric amount of research to illuminate the factors that play a role in the postediting process. Postediting was situated between translation and editing, being both similar to and different from the two tasks. Therefore, powered with data from translation process research, postediting became the subject of countless studies where different methods were utilized to see how translators worked as post-editors and what postediting involved as compared to other usual tasks of translators.

The most extensive work on the postediting process was published by Krings (2001), who compiled existing translation process data and highlighted the aspects that required attention in postediting studies. In short, Krings (2001) identified three types of effort: temporal, technical, and cognitive.

Technical effort, according to Krings's (2001) definition, refers to the changes made on the target text during postediting. Technical effort is directly measurable through different methods that are also used to compare and contrast machine translation outputs with human translation references. Such tools include Translation Edit Rate (TER), Human Translation Edit Rate (HTER), Levenshtein distance, etc. Most of these tools compare the final post-edited output (reference) with the raw output of machine translation (hypothesis) at word-level (although character-based systems [e.g. CHARCUT (Lardilleux & Lepage, 2017)] and systems which recognize synonyms [e.g. METEOR (Denkowski & Lavie, 2011)] exist) and produces a numerical output related to the amount of changes (insertions, deletions, substitutions, etc.) or the similarity between the reference and hypothesis. When used in postediting research, reference is not the pure human translation but the final post-edited output, while hypothesis remains the raw machine translation output.

Temporal effort is measured by the time spent on postediting a given text. Time per segment/sentence will directly indicate the temporal effort: the shorter the time, the lower the effort, and vice versa.

Lastly, cognitive effort relates to the mental processes involved in postediting. It is argued that direct measurement of cognitive effort is not feasible (Lacruz & Jääskeläinen, 2018) but the cognitive process can be illustrated via indirect methods from writing research or educational psychology. Keystroke logging and eye movement data are frequently utilized to elucidate the cognitive features of the postediting process, or triangulation of temporal and technical effort data is performed to illustrate the cognitive aspect.

Krings's (2001) work on effort involved in postediting has been cited countless times in the postediting literature. There are numerous studies investigating postediting effort types and testing different methods in order to determine the reliable and correct way for measurement, several of which relevant to the scope of this thesis are discussed below.

Vieira (2016) has conducted an extensive study aiming to see how different measures correlated with one another. In his multivariate analysis, the author used "subjective ratings, eye-tracking metrics, pauses and editing time" using "both professional and non-professional participants" (Vieira, 2016, p. 43). The study used ten participants, all of whom were native English speakers, and their professional experienced ranged from >0.1 year to ≤0.1 year (Vieira, 2016). The language pair studied was English-French. Vieira (2016) found that eye fixation, keyboard pauses, and temporal data (seconds per word) showed higher correlation compared to "average pause ratio, average fixation duration, pause ratio, and subjective ratings" (Vieira, 2016, p. 59). Two important findings emerge from the research: that not all pause data relate to cognitive effort and that temporal effort is worth further exploration due to the high correlation shown here (Vieira, 2016).

Koponen et al. (2012) set out to investigate whether temporal effort alone could be used to signify cognitive effort, emphasizing that all three types of effort were connected to each other and temporal effort was the most cost-effective way for measuring cognitive effort. In their study of postediting in the English-Spanish language pair, the authors drew attention to the discrepancy of postediting behaviour between participants as their understanding of the instructions given differed (Koponen et al., 2012). Indeed, prior planning and organization of what to change in the machine translation output or reviewing their corrections or moving between sentences was shown to potentially confound data obtained from keystroke logging (Koponen et al., 2012). Using a "cognitively-motivated" error typology, the easiest errors in terms of expected cognitive effort turned out to require the least time, verifying the hypothesis that postediting time could indicate cognitive effort (Koponen et al., 2012).

The above studies all used earlier systems, mostly statistical machine translation. As first neural machine translation systems came out around 2016, the literature investigating effort using neural machine translation or the differences between neural machine translation and older systems is relatively scarce. However, several studies have already been published, investigating the potential impact of neural machine translation on the aspects studied above.

Jia et al. (2019) set out from the assumption that neural machine translation was "a more promising approach to adopt than postediting of SMT" (p. 60). Citing the scarcity of articles on postediting of neural machine translation, the authors employed keystroke logging, questionnaires, and quality evaluations to compare postediting of neural machine translation with from-scratch translation from English to Chinese (Jia et al., 2019). They also used general and domain-specific texts for further comparison. Neural machine translation was found to be faster only in specialized texts, and the cognitive effort involved in neural machine translation postediting was found to be less for both domain-specific and general texts with no remarkable difference in quality between texts produced with from-scratch translation and postediting (Jia et al., 2019).

Another study by Gijon et al. (2019) compared neural machine translation postediting with the editing of translation memory matches. Although the authors do not specifically cite "cognitive effort," they still aim to quantify technical and temporal effort by way of recording editing events and time involved in either task to indicate productivity benefits (Gijon et al., 2019). Interestingly, neural machine translation required less editing, but more time compared to translation memory segments, leading to no significant differences with regard to productivity (Gijon et al., 2019). This difference was based by the authors on the similarity of translation memory matches to the source text while quality and similarity differed between neural machine translation segments (Gijon et al., 2019). Finally, perceived effort was found to affect the temporal aspect when the participants considered that neural machine translation postediting made them faster (Gijon et al., 2019).

Yamada (2019) used an older design employed in a previous statistical machine translation study to compare the new neural machine translation system with statistical machine translation in terms of cognitive, technical, and temporal effort. In English to Japanese texts, the cognitive effort showed no significant difference between neural machine translation and statistical machine translation, while the amount of editing differed significantly with neural machine translation producing better quality (Yamada, 2019). In contrast, student-translators enrolled in the study tended to correct fewer errors when postediting neural machine translation outputs, possibly due to "NMT producing human-like errors that make it more difficult for students to post-edit" (Yamada, 2019, p. 87). The author concluded that the advanced system had actually made the postediting process harder for students, as compared to his previous study, due to the complexity of errors that were produced by neural machine translation (Yamada, 2019).

### 1.3.3. Product of Postediting

Another aspect that concerns translators and language service providers, as well as clients, is quality of the product. Quality is explored in the literature from a translator's viewpoint and an end-user perspective.

The aspect of quality evaluation differs when it is applied to post-edited products. In the traditional sense, quality is evaluated as part of a translation assessment process, which is sometimes deemed a part of translation criticism (Lauscher, 2000). Literary translations are often critiqued subjectively and qualitatively with a focus on style, adequacy, and discourse. On the other hand, literary translation is outside the current scope of postediting applications. Therefore, a more practical approach is adopted when evaluating translations produced as a result of a postediting process. Usually, quality assurance typologies utilized in the industry (e.g. LISA QA model) are applied, which allows for the classification of errors in the target text in different categories (such as major/minor or accuracy, fluency, style, etc.) and produces a general score with regard to the quality of the product. What makes it different compared to translation evaluation in the traditional sense is the necessity of changes (Koponen, 2018) and attitude/behaviour of the translator towards the task or machine translation. Postediting aims for increased productivity, thus redundant changes are undesirable. There are also postediting guidelines dictating what should and should not be changed. A translator (as compared to post-editor) is not generally bound by these rules or have a specific attitude towards the task at hand. These factors mandate adjustments in the translation assessment procedure, and some examples as employed in the literature are illustrated below.

Fiederer and O'Brien (2009) set out to investigate the clarity, accuracy, and style aspects of postediting products in a setting consisting of 11 graders and 30 source sentences. The graders were asked to rate three translations and three post-edited versions of 30 sentences each on a scale of 1 to 4, with 4 representing the highest score in a given category (Fiederer & O'Brien, 2009). Accuracy was found to be higher among post-edited outputs compared to human translations, while clarity results showed no difference between the two categories (Fiederer & O'Brien, 2009). On the other hand, style in human translations was superior to postediting (Fiederer & O'Brien, 2009). This study was one of the first which explored the quality aspect, as most studies in the postediting literature are concerned with the process and not the product.

In a rater-blinded setting, Garcia (2010) explored the use of machine translation in general texts. It is worth mentioning that this study used bilinguals instead of professional translators or translation students and also explored the idea if non-translators could be successful at postediting. Two raters assessed the resulting translations (from-scratch and postediting) according to a pre-set guideline published by the Australian National Accreditation Authority for Translators and Interpreters (NAATI) (Garcia, 2010). Machine translation was chosen as the better option in 59% of the cases, with the results being insignificant for one of the evaluators (Garcia, 2010). This is an interesting finding in that the post-edited output produced by non-translators was found to be superior to from-scratch translations in more than half of the cases, suggesting the potential benefits of machine translation for non-translator bilinguals and is worth further exploration.

Another interesting study provided insight into the relation between degrees of postediting and perceptions of end-users (Egdom & Pluymaekers, 2019). The authors set out to investigate if a higher degree of postediting was necessary when a lower degree would satisfy the end-user's needs. To this end, four degrees of postediting were studied: minimal, light, moderate, and full. Text quality was found to be affected significantly by the degree of postediting applied. Moderate and full postediting applications did not demonstrate any significant differences on the informative text, while the instructive text was found positive by the end-users in two categories. Nevertheless, the authors concluded that the degree of postediting was proven to be a serious consideration on the client and the language service provider side.

In a productivity- and quality-oriented study, Depraetere et al. (2014) sought to investigate whether the productivity increase observed with machine translation postediting was accompanied by a corresponding benefit in the quality of the postediting output. For this purpose, a comparison of the postediting output and from-scratch translations were made. 181 segments were analysed on a scale of 1 to 5, a higher number indicating better quality. One professional translator (who also gave lectures on translation at postgraduate level) scored the outputs of six participants, masked to which segment was post-edited or translated from scratch. Interestingly, translations were favoured over post-edited segments for all participants, albeit with minimal differences. Although the findings did not account for a specific benefit of postediting, they still demonstrated that postediting did not significantly compromise the quality of the translation compared to from-scratch translation.

As previously stated, another issue that is widely studied is the necessity of changes made during the postediting process. How postediting instructions and/or guidelines are perceived and what minimal/maximal change means to a post-editor may affect the amount of changes and, in fact, how a resulting segment will be rated during the quality evaluation. In the above study (Depraetere et al., 2014), they triangulated human evaluation data with an automatic similarity score between the participant outputs and reference translations (assuming that the references would be of better quality) and found around 60% similarity. Other methods, using a similar logic, are also applied; automatic machine translation evaluation methods are often employed for this purpose (e.g. BLEU, TER, METEOR, etc., two of which are also utilized in the present study).

In a pilot study, Koponen and Salmi (2017) investigated this aspect of the postediting process, namely the necessity of changes and associated quality. Five students were asked to take part in an English-Finnish postediting task. The text was taken from the WMT database and consisted of news articles. It is worth noting that the participants were instructed to conduct light postediting, meaning that minimal changes would be favoured. Manual ratings were utilized for quality analysis and measuring the necessity of changes. It was determined that a quarter of all changes were related to word forms and these changes were also mostly necessary (70%) along with insertions (84%) and substitutions (67%). The majority of deletions and syntactic changes were deemed unnecessary by the evaluators. Quality was determined in terms of correctness, and around 90% of the changes were deemed correct. However, the discrepancy between the participants in terms of the amount of editing performed was noted, citing the differences in the perception of the postediting task between different types of subjects, who were translation students in this case.

### 1.3.4. Attitude towards Postediting

As part of her suggestions for necessary post-editor competences, O'Brien (2002) lists a positive attitude towards machine translation. Attitude and relative subjective aspects are later discussed by Rico and Torrejon (2012) and partly by Pym (2013), who cites a motivation for learning as one of the key skills necessary for the translator in the age of machine translation. The new developments may also instil some fears in translators, labelled as "automation anxiety" by Vieira (2018). Still, attitude towards machine translation and how it affects postediting behaviour is rarely studied in an experimental setting.

In a relevant study related to postediting training, Koponen (2015) investigated the attitudes of students towards machine translation technologies before and after a course focused on postediting. Koponen's (2015) qualitative study involved 15 students from a Finnish university enrolled in a language and translation programme. Reflective essays written after the course illuminated the attitudes of students. Koponen (2015) reported that, while most students initially had negative thoughts about expected machine translation quality and half had little to no idea about machine translation technologies, they shifted towards a more positive attitude after the course. Koponen's (2015) study suggests that insufficient knowledge about machine translation technologies may fuel adverse attitudes towards machine translation and resistance to providing postediting services.

Another study focusing on a more professional setting was conducted using two different sets of professional translators (Cadwell et al., 2018). The study used translators from a commercial language service provider and from the European Commission's Directorate General for Translation. The study aimed to investigate the reasons for adoption (or nonadoption) of machine translation among professional translators (Cadwell et al., 2018). Interestingly, translators working in the private sector thought that postediting slowed them down; in contrast, speed and productivity gains were the most common reasons to use machine translation among both groups (Cadwell et al., 2018). Terminology was the last concern among the participants as a reason not to use machine translation (Cadwell et al., 2018).

Similar to the above study, Bundgaard & Christensen (2019) explored the attitudes of professional translators towards translation memory and machine translation technologies. In a professional setting, seven translators were investigated when working on the computer-assisted translation tool, SDL Trados Studio. The translators' experience ranged between six to 23 years, and they were working as inhouse translators at a Danish translation company. The study aimed to see how translators interacted with an environment where both translation memory and machine translation results were shown. Interestingly, the outcome was that translators preferred the concordance feature (where they can search the bilingual translation memories to find meaning and context) over the other two functionalities. The authors attributed this fact to the lack of trust towards machine translation technologies. This was also shown by the fact that the participants double-checked the machine translation suggestions against the translation memory, which was also the case even when there was no apparent error in the suggestion according to the translator's opinion. This study, illuminating the underappreciated role of the concordance feature on computer-assisted translation tools, indicate the common perception towards machine translation among professionals.

Çetiner's (2019) study analysed the attitudes of Turkish translation students and how it changed after postediting training. Similar to preceding studies (Sukkhwan & Sripetpun, 2014 and Alotaibi, 2014), a positive change was observed in the students' attitudes following the postediting training in a statistically significant manner. More specifically, Çetiner (2019) reported that their trust in the accuracy that would be obtained from machine translation had increased with the training. In addition, the usual fear of the possibility of losing their job as machines would take over had reduced with the realization that machine translation still needed human intervention (Çetiner, 2019).

A common method to quantitatively measure the attitude aspect of the postediting process is by measuring perceived effort. Predicted or perceived effort can be measured before and after the experiment is conducted. Indeed, Moorkens et al. (2015) investigated the correlation between predicted and actual effort. The results did not show strong correlation, but there was still an increase in processing time as the predicted effort increased (Moorkens et al., 2015). The weak correlation can be explained by the small sample size and rating instructions for predicted effort (Moorkens et al., 2015).

A quantitative study conducted again by Koponen in 2012 involved the measurement of subjective effort in postediting, and comparisons with actual effort measures were carried out. Koponen (2012) reported the length of sentence to be an important factor in translators' perceived effort of postediting a given segment. In some cases, it was found that segments which were scored low (meaning that it would presumably require more edits) had less editing performed despite the perception of the participants.

A moderate relationship between perceptions and actual effort was also observed in a productivity study (Gaspari et al., 2014). In the experiment comparing four different language pairs (German, Dutch, English, with different directions), participants were asked for their perceptions of effort, speed, and their favourite way to work (Gaspari et al., 2014). Overall, an overwhelming bias towards translation from scratch was observed (Gaspari et al., 2014). This bias did not always correspond to an actual benefit for the preferred way of working (Gaspari et al., 2014), but the high incidence of negative attitudes towards postediting is worth noting.

All in all, no significant results were obtained when the relation between perceived and actual effort was investigated. Nevertheless, it is of importance that none of the studies directly measured the impact of a positive/negative attitude towards machine translation on actual postediting performance, which would involve a measure different than perceived effort. Inter-subject variability in the results could necessitate further research in the field with different experimental conditions that would eliminate the limitations of the studies above.

### 1.3.5. Experience and Postediting

Again, O'Brien's (2002) investigation into necessary postediting skills raise the question of whether non-translators would be good (or better) post-editors than professional translators, as the skills required by the two tasks differ. As mentioned above, studies like that of Garcia (2010) explored this very idea with promising results. However, before moving on to discovering other potential candidates for postediting tasks, it is essential to review which factors in the translation community affect the postediting performance. Several of those were discussed above with reference to relevant studies, and the remaining one is experience, which can be studied in three different ways:

- professional translators vs. subject-matter experts
- student-translators (translator trainees) vs. professional translators
- professional translators vs. educated bilinguals

(Garcia, 2010)

For the first item in the list, Temizoz (2016) used professional translators and engineers, who served as the subject-matter experts. Although engineers were bilingual with a proficient level of English and a native competence of Turkish and their jobs included translation tasks, they did not exclusively work as translators and had no translation training (Temizoz, 2016). On the other hand, professional translators were working as freelance translators and had at least three years of experience (Temizoz, 2016). A technical text was machine-translated with a public engine and presented to the participants through an online platform. The author reported no significant differences in quality between the translations produced by either group; however, terminology proved to be significantly in favour of the engineer group. Linguistic errors were fewer in the translations of professionals. It was concluded that translation training alone did not correspond to high quality when compared to engineer-translators and that postediting required language skills in addition to subject matter knowledge.

The second item is investigated to a small extent in a process- and product-oriented study conducted by Vanroy et al. (2019). More specifically, Vanroy et al. (2019) investigated how the product features affected the process of postediting, and by using two separate participant groups, students and professional translators, they also had the chance to compare these two demographics. However, the study lacks detailed examination of the differences between the groups. It is only stated that there is a clear difference in how students and professionals behave. In particular, no significant differences are detected between product and process-related data.

A more specific study exploring the differences in human translation and postediting between students and professional translators also encountered similarities between the two populations (Daems et al., 2017). Though, postediting was found to be more beneficial for translator trainees (Daems et al., 2017). The lack of significant differences was attributed to the sufficiency of the translation curriculum applied at the university where the participating students were enrolled.

Other studies also enrolled two different types of participants (Carl et al., 2011 with students and professionals; Nitzke & Oster, 2016 with professionals and "semi-professionals"); however, these studies lack comparative analyses investigating the potential variabilities between the two groups.

The small number of studies enrolling professionals alone or versus students might be attributed to the fact that researchers in the field of translation tend to use participants from their own setting rather than referring to external resources, as highlighted by Krings (2001). Krings (2001) also adds the heterogeneity of the community of translators and justifies the use of students as a way of ensuring consistency between the participants. Although Krings's work was published in 2001, the landscape of postediting research has undergone little change. A more recent review by Temizoz (2012) reveals that out of the 27 studies investigated, 7 of them used students as their participants with the remaining studies not reporting details about the population, using professionals (translators or professors/teachers), and automatic postediting systems.

### 1.3.6. Integration of Translation Memory and Postediting

Translation memory technologies provide databases that store the previous translations of the translator in order to offer them as suggestions in later, similar works. There are several similarities between translation memory and machine translation systems or editing translation memory and machine translation. Translation memory systems also work on similarity between the stored translation and the new, untranslated segment. The translation memory system then analyses the differences between the two texts and calculates a similarity score, with 100% being an exact match and anything lower being "fuzzy matches." This logic resembles earlier machine translation systems. After the translation is suggested on the computer-assisted translation tool screen, the translator can choose to utilize the previous translation and work from it instead of translating from scratch. Instead of translating, the translator (post-) "edits" the translation memory match, resembling the task of postediting.

It is not uncommon to encounter combined translation memory and machine translation suggestions. Several computer-assisted translation tools, such as Smartcat and Memsource, already provide an option to add a machine translation engine to the existing set of resources (translation memories, termbases, etc.), and the machine translation suggestion is shown along with any matches from the translation memory, permitting the translator to choose whichever suggestion is the best.

More advanced systems, like that of Lilt or SDL, employ an interactive/adaptive approach to the process, i.e., the system learns from the translator as they work. This allows simpler, more technical changes to be automatically applied without manual intervention. This is more generally called "automatic postediting" and is in earlier stages of development. However, translation memory technologies already possess part of what is promised with automatic postediting systems. "Fuzzy repair" features integrated into computer-assisted translation tools can correct matches with a higher fuzzy score; for example, numbers, abbreviations, or terms already in the termbase or translation memory can automatically be modified.

The reason for integrating translation memory and machine translation and conducting studies on the subject is due to the common attitude adopted by translators towards machine translation technologies as stated in previous sections. Combining translation memory and machine translation gives the translator a choice and helps them see the similarity between the two translation aids. There is also a correlation established between high-fuzzy matches and machine translation segments (O'Brien, 2006).

As part of a larger study delving into the postediting process, Guerberof Arenas (2008) compared productivity gains between postediting machine translation and translation memory segments. Their hypothesis that postediting machine translation and translation memory would take around the same amount of time was rejected as the processing speed was higher for machine translation segments compared to translation memory matches. It was found that the participants working more slowly took more advantage of the translation aids provided, and it was also revealed that the fastest task was the translation of new segments without any machine translation or translation memory matches. Although the limitations of the study include the inconsistent data with large gaps between minimum and maximum values resulting in higher standard deviations, these results could still have implications on whether translation memory or machine translation improves productivity in reality.

In contrast to the above findings, a more recent experiment conducted by Sánchez-Gijón et al. (2019) compared editing of neural machine translation segments with translation memory matches. While translation from scratch was found superior to the other two task types above, the authors in this study found that less editing was required for neural machine translation segments compared to translation memory segments; however, the less editing necessary in neural machine translation outputs took more time compared to the time spent on translation memory matches, leading to the researchers' conclusion that no significant productivity benefits were obtained with neural machine translation. This rather interesting and unexpected finding could be explained by the variability in the quality of neural machine translation outputs, whereas translation memory matches tend to show similarity to the project at hand as they are from older translations. It is also possible that the potential human errors that could be present in translation memory matches are not as many as those made by the machine translation engine, leading to a higher amount of temporal effort in editing of the latter.

In a mixed-design study, Teixeira (2014) compared the actual and measured performances of professional translators in three different types of tasks, namely translation, revision, and postediting. Revision and postediting tasks were randomly mixed, and revision consisted of three different levels of matches, i.e., exact and fuzzy (70-84% and 95-99%) matches. The tasks were presented in two different forms with metadata present in one and not in the other, meaning that one was a blind setting, and the participant did not know the source of the segment. Manual evaluations by two professional reviewers were utilized for quality analysis, whereas interviews were conducted to measure the perceived effort of participants. The data obtained from ten professional translators showed that from-scratch translation required the most temporal and technical effort. However, the higher amount of effort exerted when translating from scratch did not correspond to a lower amount of errors. In fact, in 70% of the cases, the errors were highest in human-translated sentences. The author explained this phenomenon by the reliance of the modern translator on translation aids, as professional translators tend to work on a computer-assisted translation tool environment with translation suggestions in one form or another. Another finding was that the presence or absence of metadata did not have a significant impact on measured performance but the interview data (perceived effort) demonstrated the prejudice among the participants towards machine translation, therefore suggesting that the source of suggestions might have had an impact on cognitive load. The translators also favoured the presence of metadata as it was more similar to the way they usually work.

Different aspects of the postediting process within the scope of this thesis were discussed above with reference to relevant studies. It was demonstrated that all studies had limitations in one way or another. The studies tended to use translation students due to logistic reasons, and in terms of quality evaluations, manual (usually two human reviewers) and automatic ratings were applied. Several paths were followed in order for the indirect measurement of cognitive effort, and cognitive effort measurements were sometimes substituted with temporal and technical effort analyses. Few studies investigated attitude towards machine translation and its effect on performance, and when a comprehensive assessment was made, quantitative measures for the comparison with the task itself were not utilized. When it comes to the Turkish literature, there is a serious scarcity of data, particularly about the postediting process. There is only one process-oriented study conducted by Temizoz in 2016, and no studies on neural machine translation with a focus on the postediting process in the Turkish language have been carried out so far. This indicates a highly unmet need for insights into the postediting process in the Turkish language (whether as source or target language), which this thesis intends to fulfil.

# CHAPTER 2
# METHODOLOGY

In this chapter, the methodological aspects of the thesis are detailed. First, the participant profile and the experiment design are described, and the computer-assisted translation tools used during the experiment are explained. The conduct of the experiment is elucidated. This chapter also presents the background of the machine translation engines that were used and/or prepared for the purpose of this thesis. Finally, the statistical analysis methods are described and justified.

Statistical analyses conducted as part of this study were carried out in R (R Core Team, 2018).

## 1.1. ETHICAL CONSIDERATIONS

The necessary ethics approval for the conduct of the experiment was obtained from the Hacettepe University Ethics Commission with the decision no. 12908312-300 dated 17 December 2019. All participants were duly informed about the purpose, design, and course of the experiment and they all provided informed consent forms before taking part in the study.

## 1.2. PARTICIPANTS

The participants enrolled for the experiment were either Master of Arts (MA) or PhD-level students studying at the department of English Translation and Interpretation at Hacettepe University. As noted in previous chapters, students are very commonly enrolled in translation process studies due to their availability and their willingness to complete the necessary tasks compared to professional translators. However, as professional experience in the translation industry would be one of the variables measured during the experiment and due to its potential impact on the results, students in the PhD programme were also invited to take part, assuming that they would have that kind of experience compared to MA students, who could have started the degree right after graduation without any professional experience.

Initially, 13 postgraduate students were planned to be enrolled. However, 1 student in the MA group had problems with installing the necessary software for taking part in the study. Therefore, the participant was excluded from the study, resulting in an experiment population consisting of 12 students.

Of the 12 participants whose data were analysed within the context of this thesis, 5 were PhD students, and the remaining 7 participants were first-year MA students (as part of a 2-year programme). All participants were actively taking classes during the fall semester when the experiment was conducted. Therefore, it was possible for the researcher to set up a controlled environment within a familiar classroom setting. The mean age of the participants was 26 (range: 22-36).

All participants had at least 1 year of professional experience in the translation industry. PhD students had 5 to 10 years of experience, while the range of experience among MA students was 1-3 years. All of the participants were translating between English and Turkish, with 2 participants additionally translating from and to French and German. The main task in the experiment was English to Turkish, thus all participants were considered eligible. No official assessment of English skills was conducted. As part of their applications to the MA or PhD programmes, the students had already demonstrated sufficient English skills. In addition, the primary language of the department for the programmes in question is English (the students were expected or had written their theses in English), and the interview part of the application had also been conducted in English, where the applicants had to demonstrate their English competence. All these points taken together, it was not deemed relevant to conduct an additional English test to grade the proficiency of participants.

## 1.3. TOOLS

### 1.3.1. Equipment

The experiment was conducted on computer environment, and due to technical deficiencies within the experiment environment, participants had to bring their own laptops. If they did not have a laptop they could bring, one would be provided by the researcher, or they would simply be excluded from the experiment. One participant in the PhD group had to use a laptop provided by the researcher as they could not bring their own. In the MA group, two laptops were provided: one to a participant who couldn't install the necessary tools on a Macintosh operating system and one to a participant who couldn't bring their own laptop. Still, the majority of the participants used their own laptop, and the familiarity with the physical features (e.g., keyboard layout) of their own computers is considered to have contributed to the ecological validity of the study. As the aim of this experiment design was to provide a working environment as close to their own as possible, this factor was a facilitating one. Apart from the participant who was originally a Mac user, all participants were familiar with the required operating system, Windows 10, and regardless of the laptop or operating system they used, as almost all computers in Turkey come with a Turkish-Q keyboard (although a less popular "F" keyboard exists designed specifically for Turkish users), the keyboard layout remained the same.

### 1.3.2. Computer-assisted Translation Tool

The computer-assisted translation tool utilized for the conduct of the experiment was SDL Trados 2017. SDL Trados 2017 was chosen because of the assumed familiarity of most translators with the tool and the fact that it is regarded as the leader among computer-assisted translation tools (*Trados Studio - Translation Software*, n.d.). The utilization of SDL Trados 2017 was also necessary because the measurement tool employed was an add-on of this computer-assisted translation tool. A newer version of SDL Trados had also been released at the time of the experiment but taking into account the shorter amount of time for which it had been available, SDL Trados 2017 was assumed to be more common among the participants compared to its newer 2019 version.

It is very rare that postediting studies investigating effort use familiar computer-assisted translation tools like SDL Trados or Smartcat (the most popular tool in the questionnaire). Instead, specific systems are designed for research purposes, such as PET (Aziz et al., 2012)or CASMACAT (Koehn, 2016), or the popular Translog II tool is utilized. However, these tools do not provide a familiar environment for the translator, although they may facilitate the necessary recording processes. For example, segmentation is one of the main functionalities of all computer-assisted translation tools and a feature to which professional translators are nowadays accustomed, yet subjects have to work on the whole document (as if on a word processor) when using Translog II. Although Translog II's sophisticated recording functionalities cannot be disregarded, the researcher set out to find alternatives for this particular experiment, which would offer a more familiar and user-friendly working environment.

### 1.3.3. Qualitivity

The measurement tool used during the experiment was an add-on of SDL Trados 2017 called Qualitivity. Qualitivity is used for measuring productivity, and the tool is intended for professional translators who wish to measure how many words they translate in a certain time period. Qualitivity also allows for the calculation of the hourly rate, therefore minimizing the related effort on the translator's part. Nevertheless, Qualitivity provides powerful measurement methods that would be useful for research purposes, including time measurement, edit distance, and keystroke logging.

The time recording and edit distance measurement features of Qualitivity were utilized for the purposes of this experiment. Qualitivity allowed the measurement of time spent per segment, which would prove valuable during the subsequent analysis. In addition, the activity report generated by Qualitivity gives edit distance and a special measure for postediting distance (in percentage). These features are useful when one desires to measure technical effort and eliminate the need of using separate tools for analysis.

Qualitivity also records keystrokes, i.e., each key press on the keyboard by the participant is recorded. However, this feature was not utilized, though the data were still recorded.

Qualitivity's in-task pausing capability made it possible for the participants to leave for breaks, e.g. when they had to visit the bathroom. When a participant wanted to take a break, they paused the plugin from the window located at the bottom of the screen, which made the measurement stop, and when they came back, the timer restarted as if the participant had never left. This feature prevented accidental measurements of idle time on segments when the participant was away from their keyboard.

## 1.4. Machine Translation Systems

### 1.4.1. Neural Machine Translation Engine

The free machine translation engine by Google, Google Translate, was used to produce the neural machine translation segments. Google Translate utilized phrase-based statistical machine translation approach until 2016, when they switched to a neural machine translation system. The neural machine translation system was initially limited to fewer languages, which incidentally included Turkish, and then expanded to the whole set of languages supported by Google Translate. Google Translate was chosen because it is easily accessible and relatively successful compared to other neural machine translation engines online (e.g. Bing Translator, Yandex Translate). The fact that Google Translate chose Turkish as one of the first languages for neural machine translation in 2016 and was the first one to do so also had an impact on the choice. Segments that were randomly chosen to be translated by the neural machine translation engine were manually entered into the web interface of Google Translate, which were then copied to the XLIFF file that was to be imported into SDL Trados 2017.

### 1.4.2. Custom Statistical Machine Translation Engine

In addition to the neural machine translation engine, a custom engine utilizing the statistical approach to machine translation was built using a free, open-source system named Moses. Moses is frequently used in postediting research (Gerlach et al., 2013; Lacruz et al., 2012; Plitt & Masselot, 2010; Toral et al., 2018 *inter alia*). *Slate* was used as an interface to facilitate the training process. The researcher also created a specific tokenizer for the Turkish language on Slate as the tokenizer available with Moses was found to perform poorly on the dataset used.

The training and deployment process of the custom engine was simple. Using Slate, the researcher simply uploaded the TMX files of the dataset, described below, and let Moses train the engine for English-Turkish. Then, using the automatic translation feature offered by SDL Trados 2017, the randomly selected segments were pre-translated using the engine created for the sole purpose of this experiment.

### 1.4.3. Dataset

The experiment dataset consisted of randomly selected sentences from an English to Turkish corpora for news texts. The said corpora were created exclusively for a machine translation project as part of the Workshop on Machine Translation in 2012 and was later published online for free use. The dataset contained of parallel news texts in the English-Turkish language pair. The news datasets published on WMT for various language pairs prove useful for postediting experiments as they are readily available and also provide reference translations approved by human translators for evaluation. Reference translations (which were of "publishable quality" in that they could be published) were useful when evaluating the quality of the translation, e.g., with BLEU.

At the end of the training process, the Slate software produced an evaluation set comprising around 2300 segments. The segments contained a source, target, and reference translation.

*Table 1. Overview of experiment set*

| | # segs | total words | mean BLEU | min | words per seg | max |
|---|---|---|---|---|---|---|
| **SMT1.0** | 20 | 249 | 1.0 | 7 | 12.45 | 29 |
| **SMT mid** | 20 | 223 | max 0.95 | 5 | 11.15 | 16 |
| **SMT low** | 20 | 172 | less than 0.15 | 4 | 8.6 | 14 |
| | | | | | | |
| **GOOGLE 1.0** | 20 | 198 | 1.0 | 5 | 9.9 | 23 |
| **GOOGLE mid** | 20 | 209 | max 0.75 | 4 | 10.45 | 18 |
| **GOOGLE low** | 20 | 176 | less than 0.15 | 4 | 8.8 | 18 |
| | | | | | | |
| **TM exact match** | 20 | 249 | N/A | 4 | 12.45 | 21 |
| **TM fuzzy match** | 20 | 202 | N/A | 8 | 10.1 | 13 |
| **TM no match** | 10 | 136 | N/A | 8 | 13.6 | 19 |
| | | | | | | |
| **Summary** | 170 | 1814 | | | 10.7 | |

As seen in the table above (Table 1), the experiment file comprised a total of 1814 words in 170 segments. The segment categories are explained below:

> SMT 1.0/GOOGLE 1.0: The segments matched the reference translation 100%, and "1.0" indicates the BLEU score obtained for these segments. Little to no effort was predicted for these segments.
>
> SMT/GOOGLE mid: The segments had a moderately good BLEU score as compared to the reference translation. Moderate to little effort was predicted for these segments.

SMT/GOOGLE low: The segments were poorly translated by the respective engine, resulting in a low BLEU score. High to moderate effort was predicted for these segments.

TM exact match: As there was no translation memory attached in the experiment set, these segments were directly taken from the reference translation set. As with the segments with a high BLEU score, these were also estimated to require little to no effort.

TM fuzzy match: The segments were artificially created from the reference translation set so as to simulate a regular working environment with an active translation memory. The reference translations were edited so that the resulting translation suggestion was not entirely correct. These segments were predicted to require moderate to little effort.

TM no match: The segments were not pre-translated, and the participant had to translate the sentence from scratch. These segments were predicted to require high to moderate effort.

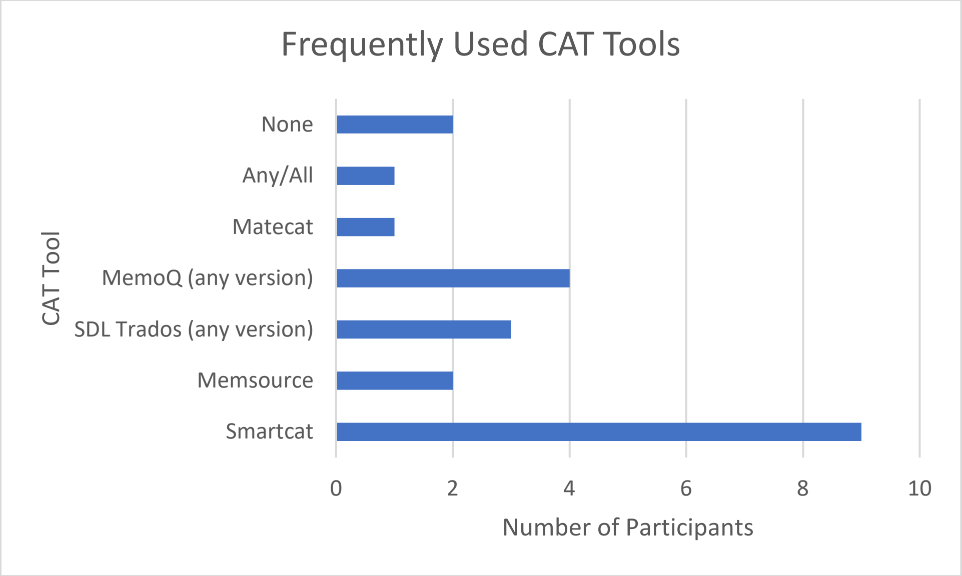
The diversity of segment categories as listed above allowed for various analyses to be conducted. Comparisons were thus possible between statistical and neural machine translation, machine translation and translation memory suggestions, and postediting and translation from scratch, among others. The main focus was to see how the main categories of "SMT," "GOOGLE" (neural machine translation) and "TM" (human translation) compared to one another. In several cases where they were not relevant, the subcategories were aggregated to create these main categories.

# CHAPTER 3
# RESULTS

## 1.1. PRE-TEST QUESTIONNAIRE

The students participating in the experiment were required to fill out a questionnaire exploring their familiarity with translation technologies (computer-assisted translation tools and machine translation), their postediting experience, and their opinion of machine translation technologies.

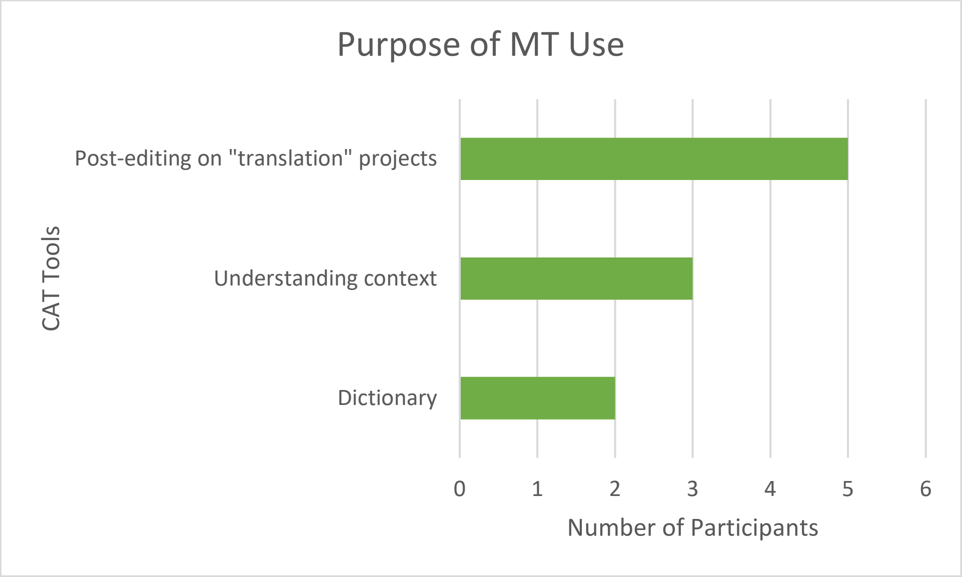
*Figure 5. Frequently used computer-assisted translation tools*



Regarding their technological competence (Figure 5), two participants (both in the PhD group) stated in the questionnaire that they did not use computer-assisted translation tools on a daily basis. All the remaining subjects were already using one or more computer-assisted translation tools and the specified tools were as follows: Smartcat (9), MemoQ (4), SDL Trados (3), Memsource (2), Matecat (1), and All [tools] (1). Considering the similarity of the computer-assisted translation tools available, the relatively less common use of SDL Trados, which was the main tool of the present study, was not considered to pose a methodological problem. SDL Trados and the main functions to be used during the task were introduced before beginning the experiment. The most frequent function required was confirming a segment, and all the tools specified above used the same shortcut, CTRL + Enter. Still, in addition to the briefing about the software at the beginning of the experiment, the participant could freely report any problems and/or refuse to take part in the experiment if they considered the software too hard to use.

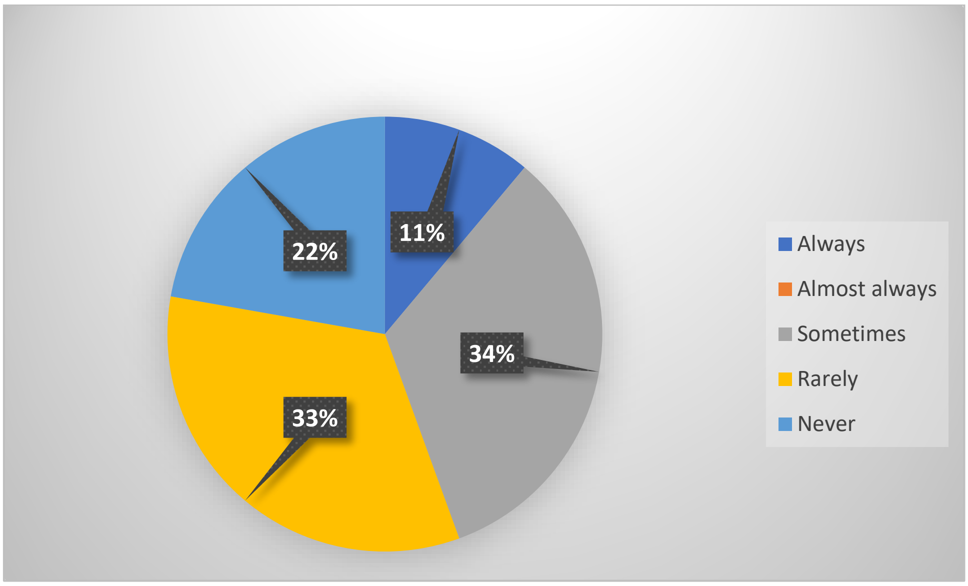
Machine translation use was surveyed before the experiment, as familiarity with different types of engine outputs could have an impact on the post-editor's performance. Regarding machine translation use, 5 participants reported no regular use of machine translation technologies, while the remaining indicated that they utilized machine translation tools for their work. Accordingly, the participants answering "yes" were asked to specify the purpose of their daily machine translation use.

*Figure 6. Purpose of machine translation use*



Interestingly, the majority of the participants indicated that they used machine translation for translation projects, i.e. when the client did not specifically instruct to use machine translation or carry out postediting. Although such use is discouraged in the industry due to several reasons, including confidentiality issues that is very common with free-to-use machine translation systems, translators can still refer to machine translation as there is virtually no way of detecting whether machine translation is used on a given document. This finding also demonstrated that, despite the lack of experience with postediting projects reported below, the participants were unknowingly conducting postediting on regular projects.

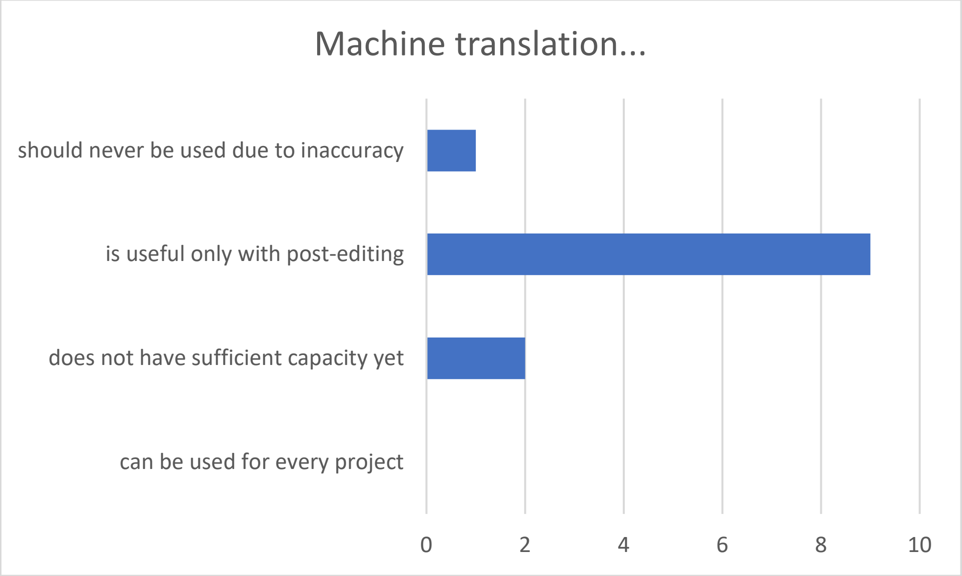
*Figure 7. Frequency of machine translation use*



Regarding the frequency of machine translation use, only 11% of the subjects reported that they always used machine translation. 33% of the subjects rarely used machine translation, while another 34% indicated that they sometimes benefited from such technologies. This picture could indicate that the sample of the study had moderate experience with machine translation.

Next, the reasons for referring to or refraining from machine translation use were investigated. The participants were asked to choose one of the four answers closest to their opinion about machine translation technologies, which would elucidate what made them use or avoid machine translation in their daily professional lives.

*Figure 8. Ideas for machine translation use*



The survey results showed that most of the participants were in favour of machine translation when it was accompanied by human postediting, while 2 subjects believed that machine translation had not achieved optimal capacity for best performance. No subject rated the highest favourable opinion for machine translation, which stated that machine translation could be used for every project, although one participant previously indicated use of machine translation for all projects. Only one participant was strongly against machine translation. Overall, the results showed that the participants enrolled in the study had a relatively positive attitude towards machine translation, with only 3 of them pointing out the deficiencies of machine translation technologies.

Finally, the participants were asked about their professional postediting experience. The scope of the question included only professional projects, where the subject was explicitly asked to carry out postediting on a source text that was pre-translated with a machine translation system. 75% of the participants did not have any postediting experience, although the earlier results above showed that they were carrying out postediting on their "translation" projects. Among those who answered yes to the aforementioned survey question, the amount of postediting projects with which they were involved were 2, 4, and "more than 10." Although postediting services are being increasingly common, it is still not surprising to see a lack of familiarity with postediting as few companies in Turkey offer postediting services, and even fewer academic institutions provide courses related to machine translation and postediting, exemplified by the online course plan for the department at which the experiment was conducted without any machine translation courses.

## 1.2. QUANTITATIVE EXPERIMENT DATA

### 1.2.1. Descriptive Statistics

The data contained a total of 2052 observations among the participants. In this section, the data are expressed as mean (standard deviation [SD]) and median (range) as applicable.

For the time measure in seconds, the mean value was 32.84 seconds (standard deviation [SD] 29.66), and the median was 23.61 (range: 0.01-625.22). For the time variable, the skewness of the data was calculated as 5.53.

For the edit distance measure calculated with the Levenshtein formula, the mean value was 20.20 (*SD* = 17.93), and the median was found as 18.00 (range [0, 121]). The skewness of the data was calculated to be 0.84.

Regarding the number of tokens, i.e. word count, in each segment, the mean word count was 10.60 words (*SD* = 4.27). The median value was 10.00 words (4, 29), and the skewness of the data was calculated to be 0.91.

In the following sections, time and edit distance data are explored; first, the distribution of the data is determined in order to designate the tests to be used during the investigation of relevant effects. Afterwards, the data are analyzed using non-parametric tests. Following the analysis of intergroup differences, the data are separately fit into a simple linear regression model in order to further explore their effects in a much more general sense.

### 1.2.2. Distribution of Data

In order to test whether parametric or non-parametric tests should be applied to the data, a density plot of raw time data in seconds and their log-transformed version was drawn. The results showed a large number of extreme values, as could be expected from a diverse sample of human translators.

Afterwards, Shapiro-Wilk normality test was carried out in order to see if the data were normally distributed. The results (statistic = 0.69) showed a p-value ($p < 0$) lower than the alpha level of 0.05, confirming that the data were significantly different from normal distribution. This resulted in the conclusion that non-parametric tests were to be used with the time data.

Edit distance data calculated with the Levenshtein formula were subjected to the same procedure as the time data. Density plots were drawn first. The plots and the Shapiro-Wilk test ($p<0.05$) showed that, as with the time data, edit distance data were also not normally distributed. As a result, the following analyses used non-parametric tests in order to detect intergroup differences and any significance in the results thereof.

### 1.2.3. Time Data

In the context of postediting effort calculation, time corresponds to temporal effort among the three categories defined by Krings (2001). For the measurement of temporal effort, the time spent on each segment was recorded in seconds by Qualitivity. In this aspect, Qualitivity is more precise than its counterparts, Inputlog and Translog. While the time is recorded from the time a 'start recording' button is pressed on the latter tools, Qualitivity records the seconds spent for each segment; therefore, a more sensitive recording procedure occurs in between each segment. In doing this, the researcher was able to differentiate between time spent on each segment type and category in a mixed XLIFF file.

Table 2 shows the summary statistics for Source categories.

*Table 2. Summary statistics of Source segments*

| Source | count | mean | SD | median | IQR |
|---|---|---|---|---|---|
| **GOOGLE** | 720 | 30.29 | 20.39 | 23.98 | 22.67 |
| **SMT** | 720 | 31.29 | 24.91 | 22.95 | 25.75 |
| **TM** | 504 | 34.86 | 42.24 | 21.32 | 34.45 |
| **Translate** | 108 | 50.69 | 33.36 | 42.35 | 39.62 |

In the table, it is demonstrated that, albeit a large standard deviation, the participants spent more time on human-translated segments versus machine-translated segments. Segments produced by Google's neural machine translation engine seems to have taken the shortest time to edit, followed by the custom statistical machine translation engine, translation memory matches, and translation from scratch.

As previously stated, Source components were further divided into Categories according to how much editing they would require. This resulted in the following Categories:

- SMT 1.0
- SMT high
- SMT low
- Google 1.0
- Google high
- Google low
- TM-fuzzy (high/low)
- TM-exact (1.0)

In Table 3, the mean time and corresponding SD values are shown in more detail for segment categories.

*Table 3. Summary statistics of Category segments*

| Category | count | mean | SD | median | IQR |
|---|---|---|---|---|---|
| **GOOGLE 1.0** | 240 | 29.37 | 21.67 | 20.36 | 26.50 |
| **GOOGLE high** | 240 | 31.86 | 21.68 | 26.20 | 25.27 |
| **GOOGLE low** | 240 | 29.64 | 17.54 | 24.86 | 15.58 |
| **SMT 1.0** | 240 | 29.00 | 25.19 | 19.03 | 27.70 |
| **SMT high** | 240 | 27.60 | 22.83 | 20.82 | 24.87 |
| **SMT low** | 240 | 37.27 | 25.62 | 31.70 | 25.85 |
| **TM-control** | 108 | 50.69 | 33.36 | 42.35 | 39.62 |

| | | | | | |
|---|---|---|---|---|---|
| **TM-exact** | 240 | 31.76 | 30.45 | 19.03 | 35.40 |
| **TM-fuzzy** | 264 | 37.68 | 50.52 | 25.07 | 29.09 |

In a more detailed view, Table 3 reinforces the more general results from Table 2, i.e., fuzzy translation memory matches generally took longer to edit than machine translation outputs.

An overview of participant time data, excluding the TM-control values (translation from scratch), is given in the tables below.

*Table 4. Overview of participants' time data*

| **Participant ID** | **Qualitivity** | **Unaccounted time** | **Total time** | **% Unaccounted** | **min:sec per segment** | **words per hour** |
|---|---|---|---|---|---|---|
| **P1** | 01:12:31 | 00:10:37 | 01:23:08 | 14.6% | 00:31 | 1210 |
| **P2** | 01:11:44 | 00:06:56 | 01:18:40 | 8.8% | 00:29 | 1279 |
| **P3** | 01:14:33 | 00:40:07 | 01:54:39 | 35.0% | 00:43 | 878 |
| **P4** | 01:02:25 | 00:38:16 | 01:40:41 | 38.0% | 00:38 | 999 |
| **P5** | 00:56:27 | 00:30:48 | 01:27:15 | 35.3% | 00:33 | 1153 |
| **P6** | 01:19:01 | 00:22:31 | 01:41:32 | 22.2% | 00:38 | 991 |
| **P7** | 01:09:03 | 00:19:21 | 01:28:24 | 21.9% | 00:33 | 1138 |
| **P8** | 00:47:10 | 00:09:35 | 00:56:44 | 16.9% | 00:21 | 1773 |
| **P9** | 00:49:14 | 00:10:19 | 00:59:33 | 17.3% | 00:22 | 1690 |
| **P10** | 00:45:38 | 00:14:56 | 01:00:34 | 24.7% | 00:23 | 1661 |
| **P11** | 01:12:56 | 00:17:01 | 01:29:57 | 18.9% | 00:34 | 1119 |
| **P12** | 01:23:31 | 00:16:42 | 01:40:13 | 16.7% | 00:38 | 1004 |
| | | | | | | |
| **maximum** | 01:23:31 | 00:40:07 | 01:54:39 | 38.0% | 00:43 | 1773 |
| **mean** | 01:05:21 | 00:19:46 | 01:25:07 | 22.5% | 00:32 | 1241 |
| **minimum** | 00:45:38 | 00:06:56 | 00:56:44 | 8.8% | 00:21 | 878 |

During the experiment, Qualitivity recorded seconds spent per segment as well as the timestamps for when the participant entered and left the respective segment. This enabled the calculation of the time spent in between segments, and this is what the

"Unaccounted" columns refer to in the segments. Total active time is indicated in the Qualitivity column, while idle time is specified in the Unaccounted column.

### 1.2.4. Edit Distance Data

Edit distance, depending on the formula used, refers to the word- or character-based differences between the baseline and final version of a segment, hence measuring the amount of editing performed by the participant. For the purposes of this experiment, Levenshtein distance was calculated for each segment included in the experiment set. Levenshtein is a character-based measure of edit distance. Postediting in Turkish may commonly include changes to a word, such as adding a suffix or prefix, without changing the word itself; therefore, a character-based measure was found more advantageous over its alternatives. When it comes to Krings's (2001) classification, edit distance corresponds to technical effort.

The tables below show the summary statistics of edit distance for segment types and segment categories.

*Table 5. Summary statistics of Source segments*

| Source | count | mean | sd | median | IQR |
|---|---|---|---|---|---|
| **GOOGLE** | 720 | 19.06 | 15.47 | 18 | 24.00 |
| **SMT** | 720 | 17.14 | 16.14 | 14 | 26.00 |
| **TM** | 504 | 22.16 | 19.85 | 21 | 38.00 |
| **Translate** | 108 | 38.96 | 22.50 | 35 | 29.25 |

As seen in the table above, human-translated translation memory segments took more technical effort compared to machine-translated segments. The custom engine outputs took the lowest effort followed by outputs produced by Google Translate.

*Table 6. Summary statistics of Category segments*

| Category | count | mean | SD | median | IQR |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **GOOGLE 1.0** | 240 | 12.342 | 13.58 | 8.5 | 23.25 |
| **GOOGLE high** | 240 | 18.446 | 16.54 | 13.5 | 26.00 |
| **GOOGLE low** | 240 | 26.404 | 12.71 | 25.0 | 14.25 |
| **SMT 1.0** | 240 | 6.183 | 11.78 | 0.0 | 7.00 |
| **SMT high** | 240 | 15.463 | 14.18 | 9.0 | 16.00 |
| **SMT low** | 240 | 29.762 | 12.68 | 28.0 | 13.25 |
| **TM-control** | 108 | 38.963 | 22.50 | 35.0 | 29.25 |
| **TM-exact** | 240 | 8.242 | 15.00 | 0.0 | 12.00 |
| **TM-fuzzy** | 264 | 34.818 | 14.53 | 35.0 | 20.00 |

When examined in more detail in Table 6, a trend can be observed where the high fuzzy and exact match segments required the lowest editing amount followed by higher fuzzy and TM-control segments. As TM-control segments needed to be translated from scratch, the highest effort seen here is very much expected as the formula would count each word addition to a value of zero. On the other hand, human-translated translation memory matches took a higher effort to edit than machine-translated segments, which will be discussed in detail in the sections below.

The table below provides an overview of edit distance data of participants.

*Table 7. Overview of participants' edit distance data*

| Participant ID | correct ed=0 | % correct ed= | incorrect ed=0 | total unchanged |
|---|---|---|---|---|
| **P1** | 22 | 47.8% | 0 | 22 |
| **P2** | 23 | 50.0% | 7 | 30 |
| **P3** | 22 | 47.8% | 4 | 26 |
| **P4** | 24 | 52.2% | 0 | 24 |
| **P5** | 30 | 65.2% | 3 | 33 |

| | | | | |
|---|---|---|---|---|
| **P6** | 18 | 39.1% | 1 | 19 |
| **P7** | 35 | 76.1% | 3 | 38 |
| **P8** | 40 | 87.0% | 4 | 44 |
| **P9** | 39 | 84.8% | 1 | 40 |
| **P10** | 42 | 91.3% | 1 | 43 |
| **P11** | 39 | 84.8% | 3 | 42 |
| **P12** | 31 | 67.4% | 4 | 35 |
| | | | | |
| **maximum** | 42 | 91.3% | 7 | |
| **average** | 30 | 65.2% | 3 | |
| **minimum** | 18 | 39.1% | 0 | 1 |

### 1.2.5. Kruskal-Wallis & Pairwise Wilcoxon Tests

#### 1.2.5.1. Time

For intergroup comparisons, the normality of the Time and Distance variables were non-normal, as indicated above. Thus, non-parametric tests were favoured for intergroup assessments. Kruskal-Wallis test was used to generally test if there were any differences between groups, which were further detailed with a Pairwise Wilcox analysis.

The hypotheses were:

$H_0$: There is no difference between the Source categories in terms of Time variable.

$H_1$: There is difference between Source categories in terms of Time variable.

*Table 8. Kruskal-Wallis test of Source segments*

| .y. | n | statistic | df | p | method |
|---|---|---|---|---|---|
| **Time** | 2052 | 57.92 | 3 | 0 | Kruskal-Wallis |

Based on the analysis of the Time variable, as the p-value was shown to be less than 0.05, Kruskal-Wallis test indicated a significant difference between Source segments at a significance level of 0.05. Therefore, the null hypothesis was rejected.

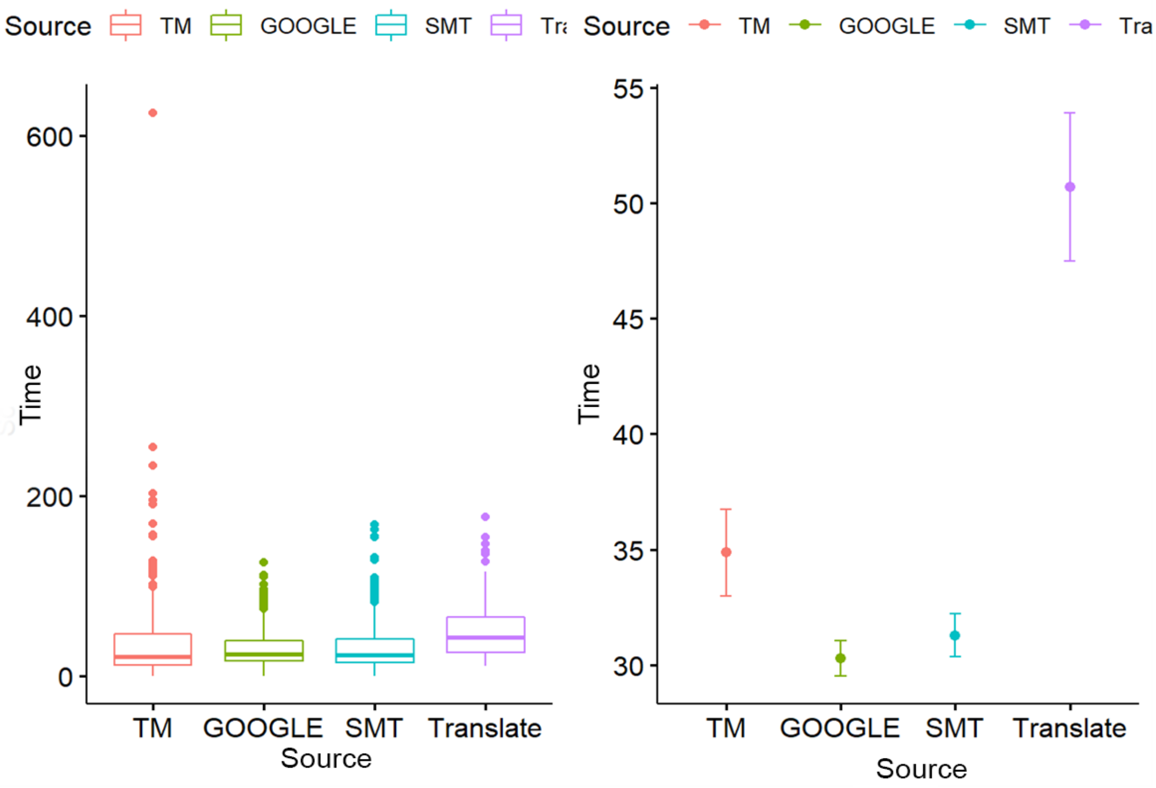
In order to explore which groups differed from one another, Pairwise Wilcoxon test was applied.

*Table 9. Pairwise Wilcoxon test of Source segments*

| .y. | group1 | group2 | n1 | n2 | statistic | p | p.adj | p.adj.signif |
|---|---|---|---|---|---|---|---|---|
| **Time** | GOOGLE | SMT | 720 | 720 | 266444 | 0.359 | 0.431 | ns |
| **Time** | GOOGLE | TM | 720 | 504 | 191012 | 0.116 | 0.174 | ns |
| **Time** | GOOGLE | Translate | 720 | 108 | 22053 | 0.000 | 0.000 | **** |
| **Time** | SMT | TM | 720 | 504 | 184865 | 0.574 | 0.574 | ns |
| **Time** | SMT | Translate | 720 | 108 | 22324 | 0.000 | 0.000 | **** |
| **Time** | TM | Translate | 504 | 108 | 15814 | 0.000 | 0.000 | **** |

Based on the findings from the Pairwise Wilcoxon analysis, all groups significantly differed from the translation from scratch (Translate) group. The difference between machine- and human-translated segment types showed no significance in terms of editing time.

*Figure 9. Box plots of Source segments for Time variable*



For a more detailed analysis, the same methodology was applied to the Categories.

*Table 10. Pairwise-Wilcoxon test of Category segments*

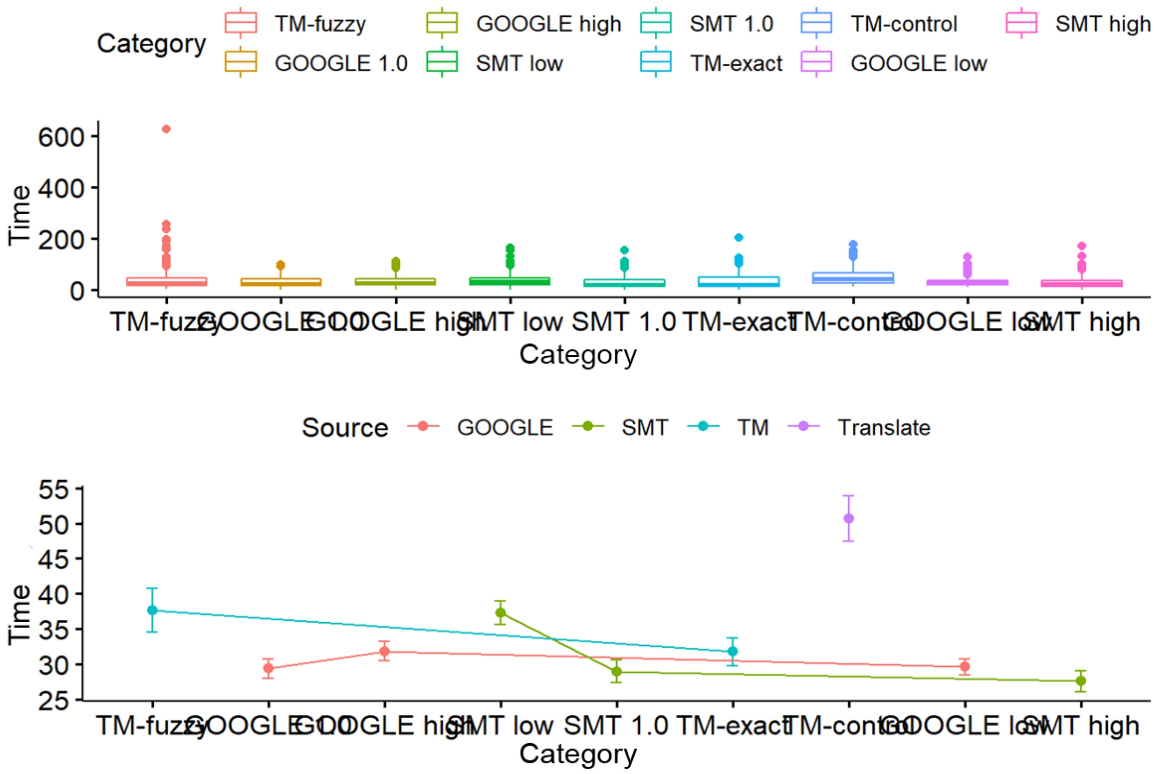
| .y. | group1 | group2 | n1 | n2 | statistic | p | p.adj | p.adj.signif |
|---|---|---|---|---|---|---|---|---|
| **Time** | GOOGLE 1.0 | GOOGLE high | 240 | 240 | 25623 | 0.0360 | 0.0510 | ns |
| **Time** | GOOGLE 1.0 | GOOGLE low | 240 | 240 | 25284 | 0.0210 | 0.0300 | • |
| **Time** | GOOGLE 1.0 | SMT 1.0 | 240 | 240 | 30636 | 0.2270 | 0.2920 | ns |
| **Time** | GOOGLE 1.0 | SMT high | 240 | 240 | 30340 | 0.3110 | 0.3730 | ns |
| **Time** | GOOGLE 1.0 | SMT low | 240 | 240 | 21930 | 0.0000 | 0.0000 | **** |
| **Time** | GOOGLE 1.0 | TM-control | 240 | 108 | 6859 | 0.0000 | 0.0000 | **** |
| **Time** | GOOGLE 1.0 | TM-exact | 240 | 240 | 30480 | 0.2690 | 0.3340 | ns |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Time** | GOOGLE 1.0 | TM-fuzzy | 240 | 264 | 28877 | 0.0860 | 0.1150 | ns |
| **Time** | GOOGLE high | GOOGLE low | 240 | 240 | 29221 | 0.7820 | 0.8550 | ns |
| **Time** | GOOGLE high | SMT 1.0 | 240 | 240 | 33775 | 0.0010 | 0.0020 | ** |
| **Time** | GOOGLE high | SMT high | 240 | 240 | 33487 | 0.0020 | 0.0040 | ** |
| **Time** | GOOGLE high | SMT low | 240 | 240 | 24619 | 0.0060 | 0.0100 | ** |
| **Time** | GOOGLE high | TM-control | 240 | 108 | 7913 | 0.0000 | 0.0000 | **** |
| **Time** | GOOGLE high | TM-exact | 240 | 240 | 33323 | 0.0030 | 0.0060 | ** |
| **Time** | GOOGLE high | TM-fuzzy | 240 | 264 | 32161 | 0.7690 | 0.8550 | ns |
| **Time** | GOOGLE low | SMT 1.0 | 240 | 240 | 34611 | 0.0001 | 0.0004 | *** |
| **Time** | GOOGLE low | SMT high | 240 | 240 | 33658 | 0.0010 | 0.0030 | ** |
| **Time** | GOOGLE low | SMT low | 240 | 240 | 23389 | 0.0004 | 0.0009 | *** |
| **Time** | GOOGLE low | TM-control | 240 | 108 | 7280 | 0.0000 | 0.0000 | **** |
| **Time** | GOOGLE low | TM-exact | 240 | 240 | 34129 | 0.0005 | 0.0010 | ** |
| **Time** | GOOGLE low | TM-fuzzy | 240 | 264 | 32042 | 0.8250 | 0.8550 | ns |
| **Time** | SMT 1.0 | SMT high | 240 | 240 | 28412 | 0.7990 | 0.8550 | ns |
| **Time** | SMT 1.0 | SMT low | 240 | 240 | 20407 | 0.0000 | 0.0000 | **** |
| **Time** | SMT 1.0 | TM-control | 240 | 108 | 6418 | 0.0000 | 0.0000 | **** |
| **Time** | SMT 1.0 | TM-exact | 240 | 240 | 28797 | 0.9980 | 0.9980 | ns |
| **Time** | SMT 1.0 | TM-fuzzy | 240 | 264 | 26886 | 0.0030 | 0.0060 | ** |
| **Time** | SMT high | SMT low | 240 | 240 | 20418 | 0.0000 | 0.0000 | **** |
| **Time** | SMT high | TM-control | 240 | 108 | 6286 | 0.0000 | 0.0000 | **** |
| **Time** | SMT high | TM-exact | 240 | 240 | 29126 | 0.8310 | 0.8550 | ns |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Time** | SMT high | TM-fuzzy | 240 | 264 | 27206 | 0.0060 | 0.0100 | ** |
| **Time** | SMT low | TM-control | 240 | 108 | 9619 | 0.0001 | 0.0004 | *** |
| **Time** | SMT low | TM-exact | 240 | 240 | 36478 | 0.0000 | 0.0000 | **** |
| **Time** | SMT low | TM-fuzzy | 240 | 264 | 36373 | 0.0040 | 0.0070 | ** |
| **Time** | TM-control | TM-exact | 108 | 240 | 18939 | 0.0000 | 0.0000 | **** |
| **Time** | TM-control | TM-fuzzy | 108 | 264 | 19679 | 0.0000 | 0.0000 | **** |
| **Time** | TM-exact | TM-fuzzy | 240 | 264 | 27230 | 0.0060 | 0.0100 | ** |

Pairwise Wilcoxon test showed a number of significant differences between the Categories. Generally, all Categories differed significantly from translation from scratch. Exact matches or segments with a BLEU score of 1.0 showed no significant difference. On the other hand, high and low fuzzy segments had a lower p-value demonstrating significance compared to other types of segments.

*Figure 10. Box plots of Category segments for Time variable*



### 1.2.5.2. Edit Distance

The same methodology was applied to test the intergroup differences based on the Distance variable.

*H*~0~: There is no difference between the Source categories in terms of Edit Distance variable.

*H*~1~: There is difference between Source categories in terms of Edit Distance variable.

*Table 11. Summary statistics of Distance variable for Source categories*

| Source | count | mean | sd | median | IQR |
|---|---|---|---|---|---|
| **GOOGLE** | 720 | 19.06 | 15.47 | 18 | 24.00 |
| **SMT** | 720 | 17.14 | 16.14 | 14 | 26.00 |

| | | | | | |
|---|---|---|---|---|---|
| **TM** | 504 | 22.16 | 19.85 | 21 | 38.00 |
| **Translate** | 108 | 38.96 | 22.50 | 35 | 29.25 |

*Table 12. Kruskal-Wallis test for Distance variable*

| .y. | n | statistic | df | p | method |
|---|---|---|---|---|---|
| **Distance** | 2052 | 99.83 | 3 | 0 | Kruskal-Wallis |

The edit distance variable was also statistically significantly difference between the groups as demonstrated by the results of the Kruskal-Wallis test. Therefore, the null hypothesis was rejected.

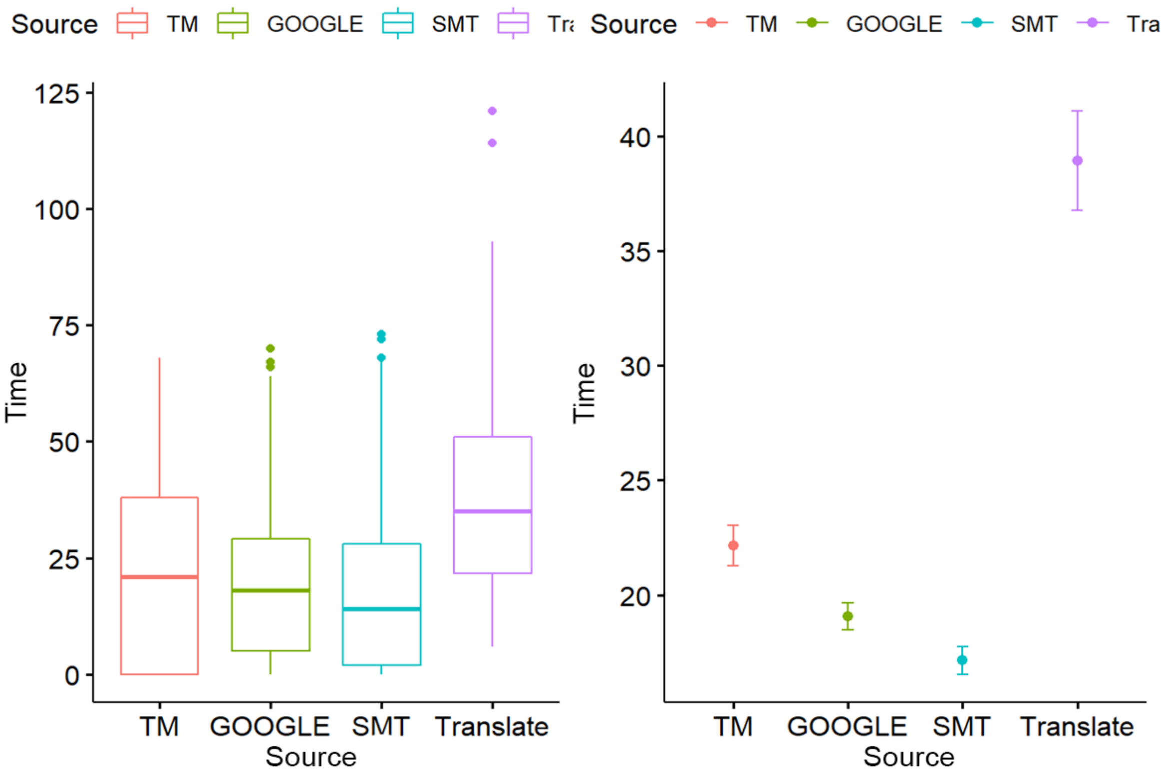
Pairwise Wilcoxon test was applied to detail the intergroup differences.

*Table 13. Pairwise-Wilcoxon test of Distance variable for Source segments*

| .y. | group1 | group2 | n1 | n2 | statistic | p | p.adj | p.adj.signif |
|---|---|---|---|---|---|---|---|---|
| **Distance** | GOOGLE | SMT | 720 | 720 | 283410 | 0.0020 | 0.003 | ** |
| **Distance** | GOOGLE | TM | 720 | 504 | 173642 | 0.1980 | 0.198 | ns |
| **Distance** | GOOGLE | Translate | 720 | 108 | 17947 | 0.0000 | 0.000 | **** |
| **Distance** | SMT | TM | 720 | 504 | 160906 | 0.0007 | 0.001 | ** |
| **Distance** | SMT | Translate | 720 | 108 | 16125 | 0.0000 | 0.000 | **** |
| **Distance** | TM | Translate | 504 | 108 | 16072 | 0.0000 | 0.000 | **** |

The Distance variable was statistically significantly different, at a significance level of 0.05, between Google and SMT, Google and Translate, SMT and TM, SMT and Translate, and TM and Translate segments.

*Figure 11. Box plots of Source segments for Distance variable*



Categories were then analysed with the Pairwise Wilcoxon test.

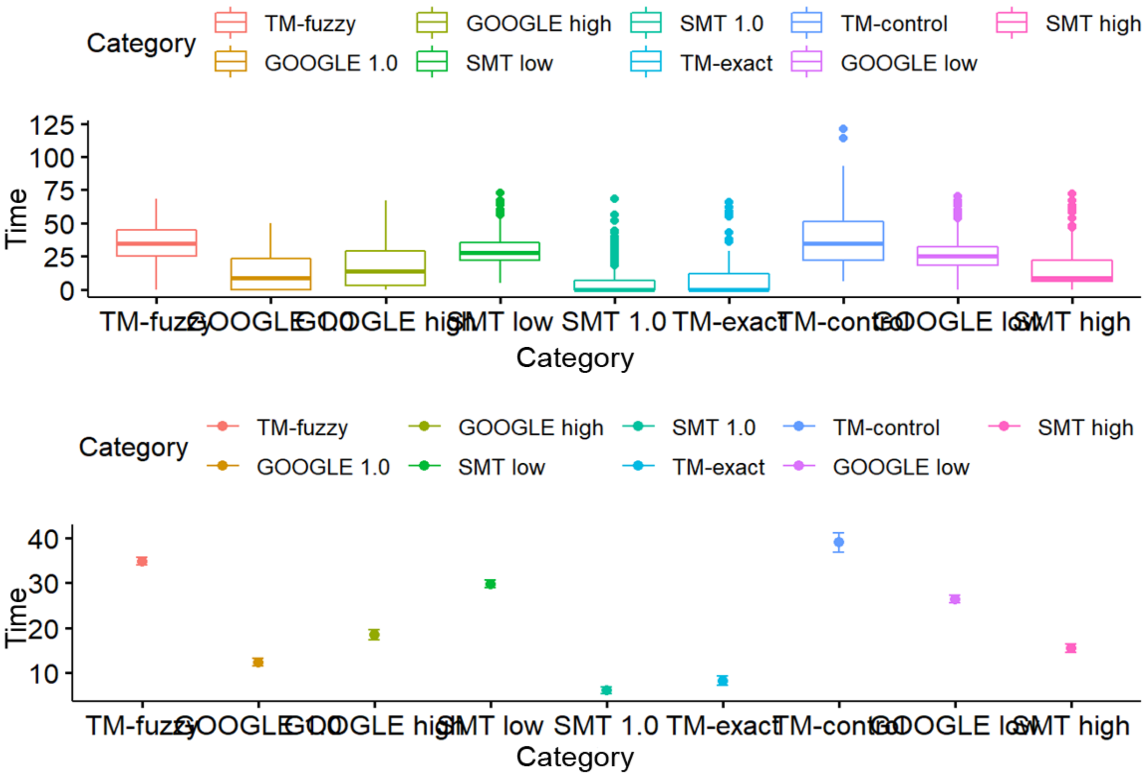
*Table 14. Pairwise-Wilcoxon test for Categories*

| .y. | group1 | group2 | n1 | n2 | statistic | p | p.adj | p.adj.signif |
|---|---|---|---|---|---|---|---|---|
| **Distance** | GOOGLE 1.0 | GOOGLE high | 240 | 240 | 21406 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE 1.0 | GOOGLE low | 240 | 240 | 12590 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE 1.0 | SMT 1.0 | 240 | 240 | 38616 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE 1.0 | SMT high | 240 | 240 | 22990 | 0.0001 | 0.0001 | *** |
| **Distance** | GOOGLE 1.0 | SMT low | 240 | 240 | 10048 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE 1.0 | TM-control | 240 | 108 | 3630 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE 1.0 | TM-exact | 240 | 240 | 36921 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE 1.0 | TM-fuzzy | 240 | 264 | 8584 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE high | GOOGLE low | 240 | 240 | 18422 | 0.0000 | 0.0000 | **** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Distance** | GOOGLE high | SMT 1.0 | 240 | 240 | 46077 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE high | SMT high | 240 | 240 | 30844 | 0.1780 | 0.1880 | ns |
| **Distance** | GOOGLE high | SMT low | 240 | 240 | 15606 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE high | TM-control | 240 | 108 | 5664 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE high | TM-exact | 240 | 240 | 44066 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE high | TM-fuzzy | 240 | 264 | 14193 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE low | SMT 1.0 | 240 | 240 | 51527 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE low | SMT high | 240 | 240 | 43708 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE low | SMT low | 240 | 240 | 23996 | 0.0020 | 0.0020 | ** |
| **Distance** | GOOGLE low | TM-control | 240 | 108 | 8652 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE low | TM-exact | 240 | 240 | 49676 | 0.0000 | 0.0000 | **** |
| **Distance** | GOOGLE low | TM-fuzzy | 240 | 264 | 20203 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT 1.0 | SMT high | 240 | 240 | 11858 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT 1.0 | SMT low | 240 | 240 | 4856 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT 1.0 | TM-control | 240 | 108 | 1754 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT 1.0 | TM-exact | 240 | 240 | 27511 | 0.3300 | 0.3390 | ns |
| **Distance** | SMT 1.0 | TM-fuzzy | 240 | 264 | 4348 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT high | SMT low | 240 | 240 | 11478 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT high | TM-control | 240 | 108 | 4118 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT high | TM-exact | 240 | 240 | 43569 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT high | TM-fuzzy | 240 | 264 | 10102 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT low | TM-control | 240 | 108 | 10252 | 0.0020 | 0.0020 | ** |
| **Distance** | SMT low | TM-exact | 240 | 240 | 51002 | 0.0000 | 0.0000 | **** |
| **Distance** | SMT low | TM-fuzzy | 240 | 264 | 24373 | 0.0000 | 0.0000 | **** |
| **Distance** | TM-control | TM-exact | 108 | 240 | 23406 | 0.0000 | 0.0000 | **** |
| **Distance** | TM-control | TM-fuzzy | 108 | 264 | 14955 | 0.4580 | 0.4580 | ns |
| **Distance** | TM-exact | TM-fuzzy | 240 | 264 | 6294 | 0.0000 | 0.0000 | **** |

Almost all categories significantly differed from one another when the segments were analysed according to Categories, except for Google-high<>SMT-high, SMT-1.0<>TM-exact, and TM-control<>TM-fuzzy segments.

*Figure 12. Box plots of Category segments for Distance variable*



### 1.2.6. Linear Regression Models

In order to model the relationship between Time and Distance variables and different groups and sub-groups, four different linear regression models were fitted.

In models 1 ($R^2$ = 0.024) and 2 ($R^2$ = 0.034), the effect of time on Source and Category segments were analysed, respectively, while in models 3 ($R^2$ = 0.072) and 4 ($R^2$ = 0.346), the effect of distance on the two categories of segments were investigated.

For Source segments, Google, TM, and Translate were found to significantly affect the temporal effort of the participants ($p<0.01$ for all). For Categories, Google 1.0, SMT-low, TM-control, and TM-fuzzy categories statistically significantly affected the time variable ($p<0.01$ for all).

When the relationship between edit distance and Source segments were analysed, all Source segments were found to have a significant impact on edit distance ($p<0.01$ for all). All Categories were found to significantly affect the edit distance ($p<0.05$ for SMT-high, $p<0.01$ for the rest).

## 1.3. POST-EXPERIMENT QUESTIONNAIRE

The questionnaire administered to the participants immediately after the experiment has ended aimed to obtain subjective data about the general attitude of the participants as well as their conception regarding the experiment. The questionnaire was prepared and applied in Turkish, and it contained questions from Paas et al.'s (2003) cognitive effort scale, aiming to provide another albeit subjective measure of the effort exerted during the experiment.

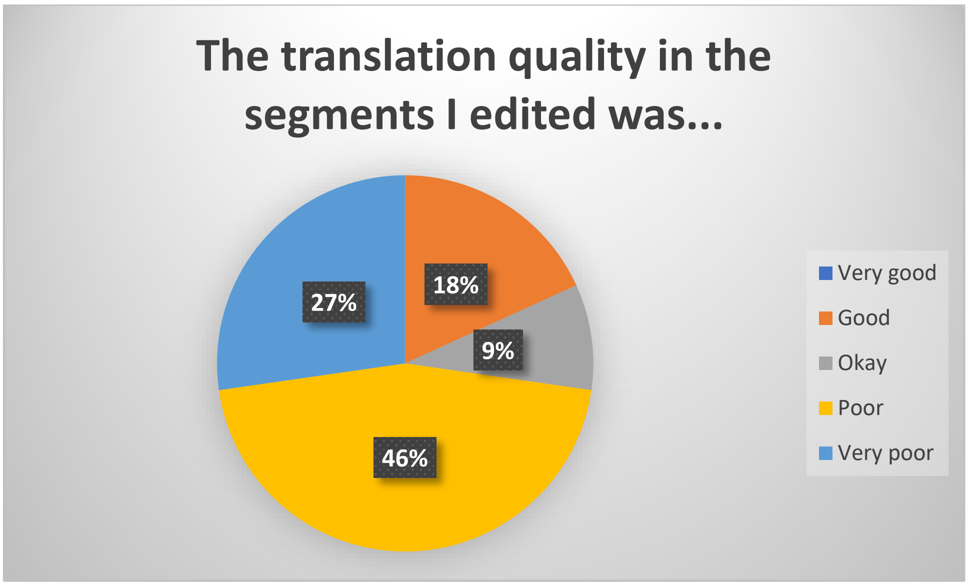
### 1.3.1. Source of Segments

The participants indicated whether they were able to understand, without any information given, where the segments came from (translation memory match, statistical machine translation, or neural machine translation). 10 participants answered “Yes” on the questionnaire, claiming that they could recognize the source of the segments. In their opinion, they were able to differentiate between machine translation and translation memory segments.

### 1.3.2. Translation Quality

The participants were asked to rate the general translation quality in the segments they performed "editing" or "postediting" on a scale of 1 to 5, where 1 = very good and 5 = very poor. The below figure includes data from 11 participants; 1 participant wrote in their own option instead of using the scale provided, which rendered their answer unusable.

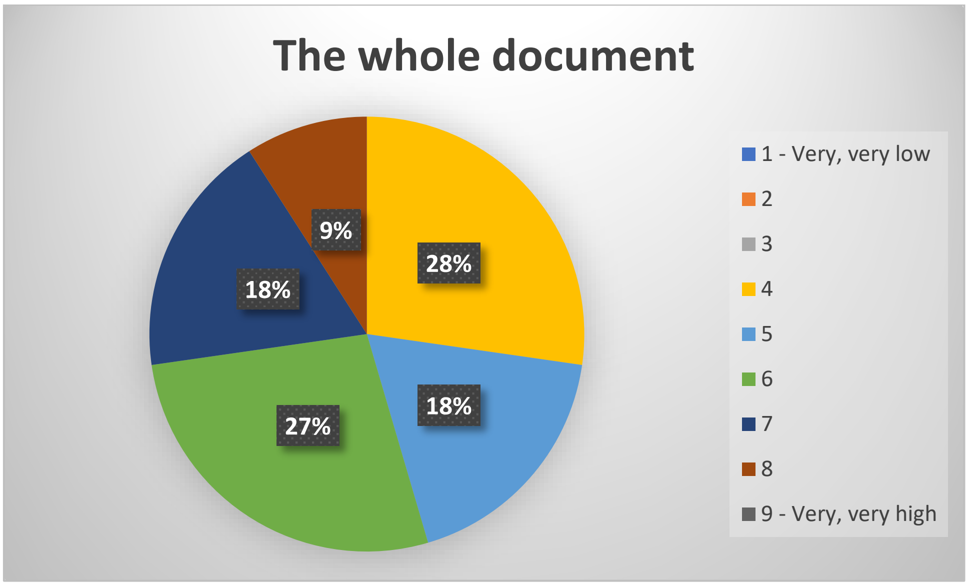
*Figure 13. Translation quality*



### 1.3.3. Subjective Effort Measurement

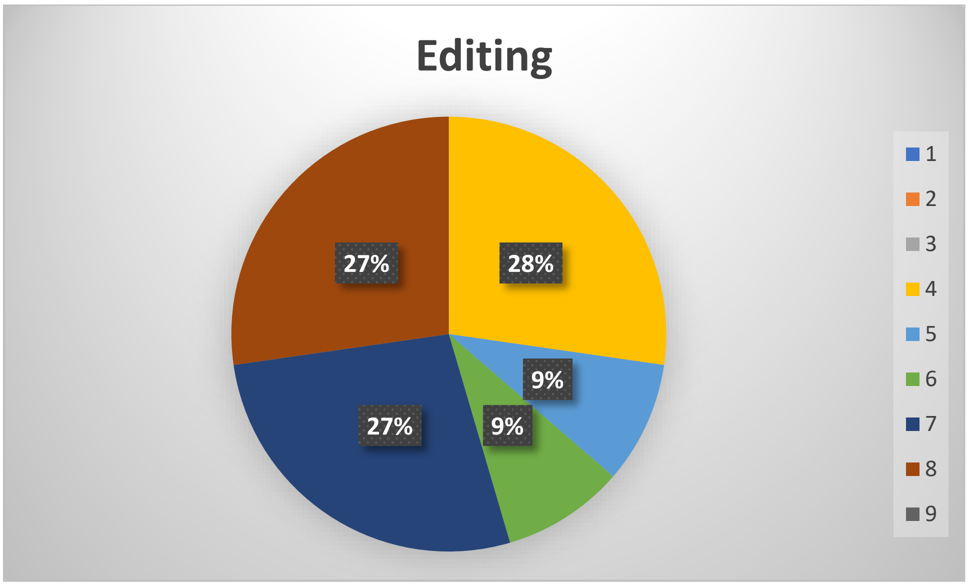
The participants were asked to rate on their own how much mental effort they exerted during various parts of the experiment: the whole document, the editing part, and the translation-from-scratch part.

*Figure 14. Subjective effort measurement: whole document*



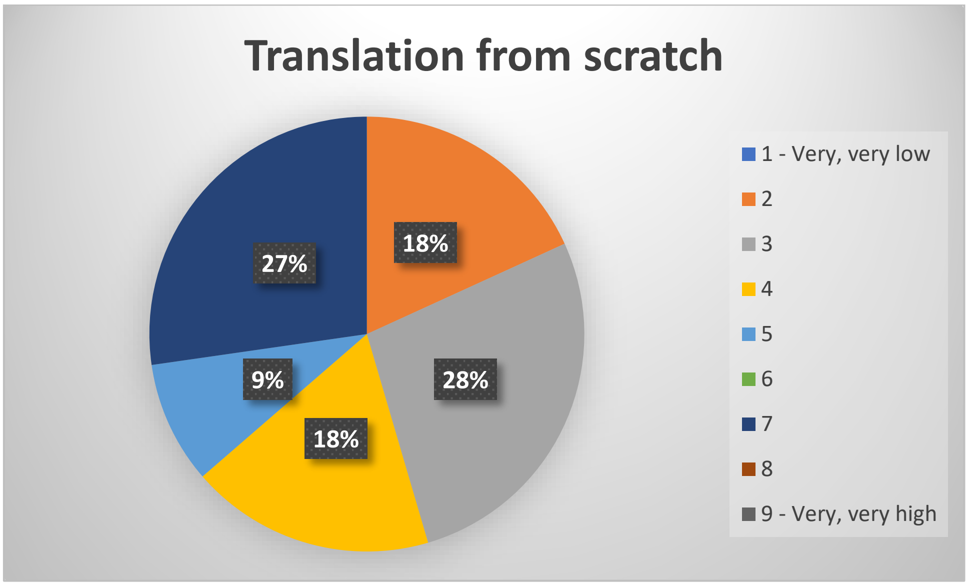
As Figure 14 demonstrates, the participants' perception of their effort regarding the whole document varies. While 18% of the participants reported a low mental effort for the whole document, none marked 9, meaning the highest effort. 9% indicated a close-to-highest effort, marking 8 on the scale, followed by 18% and 27% marking 7 and 6, respectively.

*Figure 15. Subjective effort measurement: editing*



This question relates to the whole editing experience of the participants as they were blinded to the source of the segments. While none of the participants indicated a "very, very high" effort regarding editing (regardless of human or machine translation), a total of 63% reported a generally high effort (27% for 8, 27% for 7, and 9% for 6).

*Figure 16. Subjective effort measurement: translation from scratch*



28% of the participants marked the highest effort on the scale for translation from scratch. When examined together, a total of 55% regarded the translation part of the segment as requiring a somewhat higher effort, while the rest of the participants (45%) could be judged to have regarded the task to be easier compared to the other parts of the experiment.

# CHAPTER 4
# DISCUSSION

The findings of the experiment are discussed under two subheadings, temporal effort (time) and technical effort (edit distance), with separate analyses for the two subcategories of segments.

## 1.1. TEMPORAL EFFORT

Temporal effort is a critical aspect of the postediting process in a fast-moving translation industry. How long a translator takes to complete a postediting task at hand would substantially affect the decision of the language service provider to favour machine translation over human translation. It has also been suggested in a previous study (Koponen et al., 2012) that temporal effort could be indicative of cognitive effort to some extent.

In our experiment, temporal effort was measured based on time spent on each segment as recorded by the Qualitivity tool. Segment enter and exit times were calculated by the tool, which were then used to calculate how long a participant spent on each segment. Segment enter and exit times also allowed for the calculation of "unaccounted" idle time between the segments.

The segments were categorized into two, Source and Category segments. The Source segments were:

> Google, where the participant had to edit outputs produced by Google Translate
>
> SMT, where the participant had to edit outputs produced by the custom machine translation engine created for the purposes of this thesis
>
> TM, where the participant had to edit fuzzy matches from a translation memory consisting of translations produced by humans, and
>
> Translation, where the task was to translate from scratch without any aid.

The Category segments were more detailed, and these segments were categorized according to their quality as evaluated relative to reference human translations. If a Category segment was exactly the same as the reference translation, a score of 1.0 was assigned. “High” and “low” fuzzy categories indicated that the segments highly or poorly resembled the reference translation, respectively.

Little to no difference was found between the custom engine and Google Translate. Kruskal-Wallis test for the comparison of the segments from these two sources indeed demonstrated that there was no significant difference ($p>0.05$).

In general, human-translated translation memory matches took longer time to edit compared to machine-translated segments without any statistical significance ($p>0.05$).

The time required for translations from scratch differed statistically significantly ($p<0.05$) from all other Source categories. Translation tends to take more time than editing in general, so this outcome could be regarded as expected.

One interesting finding here is that the participants spent more time on human translations instead of machine translations. The human translations included in the experiment were taken from a set of reference translations and edited to make them look like fuzzy matches. To the best of our knowledge, this finding is not reflected in any other publication in the literature. Considering the corpus that was used for this experiment was regarded as publication ready, meaning the quality was perfect or near-perfect, there is no explanation for such a trend.

Although non-significant, the difference between the custom statistical machine translation engine and Google’s neural machine translation engine indicates that, although the custom engine was trained on a specific dataset of news articles used for the experiment, statistical machine translation engines still lack in their limited understanding of language in that neural machine translation engines tend to provide more accurate and fluent outputs that require less editing time thanks to its novel technology mimicking the human brain.

When it came to the time data for Categories, similar to the results above, it was shown that TM-fuzzy segments took a longer time to edit than most of the Source categories and also had the most outlier values.

Following TM-fuzzy segments were the SMT-low fuzzy segments in terms of the highest time spent.

It is also interesting to note that TM-exact segments had a similar amount of editing time compared to Google-high fuzzy segments.

All in all, these results elucidate a complex postediting process, where the participants spent more time on human translations and not on machine translations as would be expected from them. In addition, the custom engine specifically trained for this experiment on a set of similar texts can be regarded as performing poorly when compared to the engine of Google Translate.

## 1.2. TECHNICAL EFFORT

Edit distance refers to the technical effort aspect in Krings's (2001) classification and directly illustrates how much editing has been carried out on a particular segment. Technical effort has not been attributed to cognitive effort previously but when combined with temporal effort, the amount of editing performed may indicate the extent of the cognitive effort exerted by the translator.

Technical effort is important in elucidating the complex process of postediting. Machine translation in a professional setting is expected to help the translator, thus the translator has to perform editing to a lesser extent when compared to TM matches or translation from scratch. Otherwise, the impact of the machine translation systems on the productivity of the translator may be regarded as poor.

As discussed in previous sections, edit distance in this thesis was calculated based on the Levenshtein formula, which calculates the additions, deletions, and substitutions between an original and a reference segment on a character basis. Since postediting in Turkish would involve changing the prefixes/suffixes of words in a sentence, a character-based approach was deemed suitable for the purposes of this thesis.

In terms of Source categories, the results demonstrated that there was a significant difference ($p<0.05$) between all groups except for Google and TM segments.

Translation memory segments were also found to require more technical effort compared to machine translation segments.

Although the time data discussed in the previous section showed that Google segments took less time to edit than SMT segments, but edit data indicate that more changes were applied in GOOGLE segments than in SMT segments. The difference in time spent between Google and SMT segments was not significant, however the difference in edit distance between these two segment sources is statistically significant ($p<0.05$).

Again, the segments that were translated from scratch required the highest technical effort but this outcome is predictable since the formula used would compare the changes to an empty segment and would consider every word an addition, resulting in an increased score.

For Categories, aside from three pairs (Google-high<>SMT-high, SMT 1.0<>TM-exact, and TM-control<>TM-fuzzy), all Category pairs statistically significantly differed from one another.

Low fuzzy machine translation segments (Google-low, SMT-low) and TM-fuzzy segments required more changes compared to other Categories.

This detailed examination of Categories supports the results for Source categories discussed above. Interestingly enough, the findings here suggest that TM-fuzzy segments required almost as much technical effort as TM-control segments that needed to be translated from scratch.

Overall, the findings indicate that, when the participants were blinded to where the segments came from, i.e., when they didn’t know which one was machine translation and which one was human translation, they still considered the human translated segments as requiring more editing compared to machine translated segments. This contrasts with the post-experiment questionnaire findings in which the participants rated the postediting process and the quality of the machine translation segments as poor. Although some participants indicated that they were able to recognize the source of the segments, it is very possible based on these findings that they might have mistaken human translated segments for machine translation.

In line with the results above, the linear models fitted separately for time and edit distance data revealed that, for Source segments, Google, TM, and Translate segments had a statistically significant effect on time spent. When it came to edit distance, all

Source categories, i.e., Google, TM, SMT, and Translate statistically significantly affected the technical effort exerted.

When Categories were examined under the linear model, Google 1.0, SMT-low, TM-control, and TM-fuzzy categories had a statistically significant effect on time while all categories significantly affected the amount of editing.

The present thesis aimed to evaluate and compare the temporal and technical effort associated with editing machine- and human-translated segments in a language pair that has been relatively less explored.

There is a lack of studies involving neural machine translation, statistical machine translation, and translation memory at the same time in the literature. Vieira's 2016 study indicated that temporal effort was suggestive of cognitive effort as the usual methods employed in evaluating this type of effort was not always indicative of cognitive effort. In addition, Koponen (2012) stated that methods like keystroke logging or eye-tracking could confound the data used to measure cognitive effort. Thus, temporal and technical effort could be more promising in assessing the effect of the postediting process. It can also be said that temporal and technical effort are easier to measure and more practical for the industry in general.

Similar studies have found usually non-significant differences between postediting statistical and neural machine translation outputs. In one study (Jia et al., 2019), neural machine translation was found to be edited faster compared to statistical machine translation outputs. Gijon et al. (2019) also found that neural machine translation required less technical effort but conflictingly, more time.

In the present study, a similar trend towards conflicting results is seen. When time data are examined, despite the non-significant differences, Google Translate outputs are found to be faster to edit than statistical machine translation and translation memory segments. On the other hand, the edit distance data obtained demonstrate that more changes might have been implemented in neural machine translation outputs compared to the other two types of segments.

Yamada's 2019 study found that student-translators tended to correct less errors when working with neural machine translation because of neural machine translation's ability to produce more fluent outputs with less obvious errors. This is not reflected in the current results as the edit distance seems to be higher with neural machine translation compared to statistical machine translation.

Another interesting trend seen in the findings of the current experiment is that student translators both spent more time and edited more when working with human-translated translation memory fuzzy matches. It's worth noting that these fuzzies were created artificially, meaning that reference translations were edited with the addition of errors to make them look like fuzzy matches. Still, the amount of errors added was limited, and the texts used were of publication-ready quality. Therefore, this outcome is an unexpected result that is worth further exploration.

The post-experiment questionnaire revealed that some of the participants were able to recognize the source of the segments. When this finding is interpreted with the quantitative data obtained, it seems very possible that some participants thought that they were working on machine translation-produced segments when they were actually editing human translations.

Studies in the literature have so far used different environments for such experiments. These environments rarely reflect the actual working environment of a professional translator and could easily confound the findings. The present study used a state-of-the-art and popular computer-assisted translation tool with the participants' own equipment, meaning that the participants worked on the experiment file as any other job they might have gotten from language service providers, eliminating any potential interference and ergonomic problems associated with unfamiliar software/environment. Accordingly, Läubli et al. (2013) argues that the postediting process should be assessed in a realistic environment, which the present study has provided.

The perceived effort data varied substantially among the participants. The student translators all rated the translation task as the hardest part of the experiment, while their ratings for the editing part showed a moderate-to-high level of perceived effort. All in all, the common method of measuring perceived effort could be regarded as unreliable and might easily be confounded by the personal attitudes of the participants.

# CONCLUSION

This thesis aimed to elucidate the complex process of machine translation postediting by examining the temporal and technical efforts exerted by a group of student translators in a relatively less explored language pair that is English-Turkish.

Participants were asked to first complete a pre-experiment questionnaire and then complete a postediting/translation task in a common environment used by professional translators. During the experiment, the time spent on each segment was recorded for the purposes of measuring temporal effort. Technical effort was then calculated following the experiment using the Levenshtein formula. After the experiment was completed, the participants completed a post-experiment questionnaire that explored their general perceptions of the task, the machine translation engines used, and their exerted effort.

The present thesis used a blinded approach in that the participants did not know which translation came from where. Hence, they were unaware of the sources of the segments, which precluded the impact of their personal attitudes about machine translation on the task itself.

In this limited population consisting of MA- or PhD-level participants, the majority worked as a professional translator in the industry but less than half of the participants used machine translation regularly. Most supported the use of machine translation under the supervision of human translators.

For the purposes of the experiment, in which a dataset consisting of news texts were used, a custom statistical machine translation was trained on a similar corpus. For neural machine translation, the publicly available neural machine translation engine Google Translate was used. Translation memory matches were extracted from the aforementioned dataset and made to look like fuzzy matches. Finally, there were a few segments left empty in order for the participants to translate from scratch.

The experiment results showed that, for the time variable, there were significant differences between machine translation/translation memory segments and translation from scratch. The time spent for editing machine translation outputs did not differ significantly between the different types of engines. However, participants were found to spend less time on neural machine translation than in statistical machine translation segments, and interestingly, they spent more time editing human-translated fuzzy matches.

For edit distance, significant differences were observed for all segment groups. Conflicting with the above results, Google Translate outputs were edited more intensely than statistical machine translation segments; however, the human-translated translation memory fuzzy matches were edited the most among the segments that required editing.

The data were then fitted into two separate linear models for time and edit distance. Neural machine translation, translation memory, and translation from scratch were found to significantly affect time, but there was no significance observed for statistical machine translation. For edit distance, however, all groups of segments significantly affected the technical effort exerted.

The post-experiment questionnaire revealed that the participants thought that they could recognize which segments were machine translation outputs and which segments were human translations. However, the results discussed above suggest that they might have mistaken human translations for machine translation. The perceived effort measured with a simple scale showed that the majority rated the task of translation from scratch as the hardest, and the editing task was of moderate-to-high difficulty.

It can be concluded from the post-experiment questionnaire findings that perceived effort and actual temporal and technical effort are not associated with one another. In addition, the participants' spending more time on human translation could indicate the different perceptions regarding quality. It could also be said that neural machine translation and statistical machine translation might have produced more fluent and accurate outputs compared to the human translations that were extracted from a dataset of publication-ready quality.

Despite the non-significant findings, participants generally spent less time and did less editing on statistical machine translation and neural machine translation outputs compared to translation memory matches and translations from scratch. Therefore, it can also be concluded that machine translation systems increase the productivity of the human translator with regard to speed and technical effort.

The present study used a familiar professional working environment with the participants' own equipment in order to create a realistic experiment in contrast to the majority of the studies in the literature employing research-focused tools that did not resemble a typical translation environment of a professional translator.

Still, there are a few limitations to this study that are worth exploring in the future. This experiment used a corpus of news texts because of the limited number of publicly available datasets for the respective language pair. As postediting performance could easily vary according to the type of text, different text types should be preferred in future studies in order to test the impact of machine translation technologies on post-editor performance.

There is a number of different engines available for the English-Turkish language pair. This study used a common one, Google Translate, but other engines should also be evaluated further in order to determine if the most popular engine is actually the most successful one in this particular language pair.

Cognitive effort was not directly measured in this experiment but was rather evaluated as an extension of temporal and technical effort. In the literature, cognitive effort is usually measured with methods such as keystroke logging or eyetracking. There is currently no study in the Turkish translation literature employing such methods. It's also worth noting that methods such as eyetracking include the use of equipment that might interfere with the working environment of the translator.

Perceived effort was measured using a simple scale that has previously been used in similar studies. However, this scale might prove inadequate in measuring the actual perception of the participant regarding the effort they exerted. Therefore, new scales might be developed and tested, or different qualitative methods such as interviews or focus groups could be employed to collect data about perceived effort in postediting tasks.

All in all, the present thesis has some useful implications for the industry. The questionnaire results suggest that translators expect to become frustrated with machine translation postediting tasks. In addition, the participants enrolled in this experiment mistook the human translations for machine translation, probably depending on the level of quality. While they rated their editing effort to be higher, the quantitative data indicated the opposite. Indeed, machine translated segments increased the translators' speed compared to editing translation memory matches or translation from scratch.

The increase in speed with machine translation could implore language service providers to evaluate the potential integration of the technology into their usual workflow. However, while doing that, they should take into consideration the general attitude of translators towards machine translation. The main reason for the negative opinions about machine translation among translators is the fear of being replaced as well as getting paid less in an already-underpaid profession and the assumption that the quality will be poor. Therefore, stakeholders in the language industry should strive to communicate with their translators, handling any possible doubts about the impact of machine translation on their work.

# BIBLIOGRAPHY

Alotaibi, H. (2014). Teaching CAT Tools to Translation Students: an Examination of Their Expectations and Attitudes. *AWEJ*, *3*, 65–74.

Anastasiou, D. (2008). *Idioms in example-based machine translation*. https://www.academia.edu/30082514/Idioms_in_example_based_machine_translation

Aranberri, N. (2017). What Do Professional Translators Do when Post-Editing for the First Time? First Insight into the Spanish-Basque Language Pair. *HERMES - Journal of Language and Communication in Business*, *56*, 89–110. https://doi.org/10.7146/hjlcb.v0i56.97235

Aziz, W., Castilho, S., & Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3982–3987. http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf

Bundgaard, K., & Christensen, T. P. (2019). Is the concordance feature the new black? A workplace study of translators' interaction with translation resources while post-editing TM and MT matches. *The Journal of Specialised Translation*, *31*, 14–37.

Cadwell, P., O'Brien, S., & Teixeira, C. S. C. (2018). Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives: Studies in Translatology*, *26*(3), 301–321. https://doi.org/10.1080/0907676X.2017.1337210

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. L. (2011). The Process of Post-Editing: a Pilot Study. *8th International NLPCS Workshop*, *August*, 131–142.

*CAT Tools | Software Comparison Tool*. (n.d.). Retrieved May 18, 2022, from https://www.proz.com/software-comparison-tool/cat/cat_tools/2

Çetiner, C. (2019). *Makine çevirisi sonrası düzeltme işleminin çeviri öğrencilerinin tutum ve çeviri performanslarına etkisi* [Unpublished PhD Thesis, Gazi University]. https://doi.org/10.29000/rumelide.649333

Chemvura, T. (2017). *LARMAS - Language Resource Management System*. https://doi.org/10.13140/RG.2.2.34784.38405

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*, *62*(2), 245. https://doi.org/10.7202/1041023ar

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 85–91. https://doi.org/10.1080/00288306.2004.9515087

Depraetere, I., Sutter, N. de, & Tezcan, A. (2014). *Post-edited quality, post-editing behaviour and human evaluation: a case study*. 78-108. https://hal.archives-ouvertes.fr/halshs-01060447

Diño, G. (2018). *Google, Facebook, Amazon: Neural Machine Translation Just Had Its Busiest Month Ever | Slator*. Slator. https://slator.com/technology/google-facebook-amazon-neural-machine-translation-just-had-its-busiest-month-ever/

DuPont, Q. (2018). *The Cryptological Origins of Machine Translation*. AModern. https://amodern.net/article/cryptological-origins-machine-translation/

Egdom, G. van, & Pluymaekers, M. (2019). Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end users. *The Journal of Specialised Translation*, *31*, 158–176.

Farooq, U. (2018, March 21). *Neural Machine Translation with Code*. Medium. https://medium.com/@umerfarooq_26378/neural-machine-translation-with-code-68c425044bbd

Fiederer, R., & O'Brien, S. (2009). Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation*, *11*, 52–74.

Flanagan, M., & Christensen, T. P. (2014). Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer*, *8*(2), 257–275. https://doi.org/10.1080/1750399X.2014.936111

Folaron, D. A. (2010). Networking and volunteer translators. In *Handbook of Translation Studies* (pp. 231–234). John Benjamins Publishing Company. https://doi.org/10.1075/hts.1.net1

Garcia, I. (2010). Is machine translation ready yet? *Target*, *22*(1), 7–21. https://doi.org/10.1075/target.22.1.02gar

Gaspari, F., Toral, A., Kumar Naskar, S., Groves, D., & Way, A. (2014). Perception vs Reality: Measuring Machine Translation Post-Editing Productivity. *AMTA 2014*.

Gerlach, J., Porro Rodriguez, V., Bouillon, P., & Lehmann, S. (2013). Combining pre-editing and post-editing to improve SMT of user-generated content. *Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice*, *2*(1), 45–53.

Gross, A. (1992). Limitations of computers as translation tools. In J. Newton (Ed.), *Computers in Translation: A Practical Appraisal*. Routledge.

Guerberof Arenas, A. (2008). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus The International Journal of Localisation*, *7*(1), 11–21.

Hutchins, J. (1999). *Milestones in machine translation — No.6: Bar-Hillel and the nonfeasibility of FAHQT*. http://www.mt-archive.info/Weaver-1949.pdf

Hutchins, W. J. (1986). *Machine translation: past, present, future*. Ellis Horwood ; Halsted Press.

Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *Journal of Specialised Translation*, *31*, 60–86.

Koehn, P. (2016). *CASMACAT: Advanced Computer Aided Translation Beyond Postediting*. https://ec.europa.eu/info/sites/info/files/tef2016_koehn_en.pdf

Koponen, M. (2012). *Comparing human perceptions of post-editing effort with post-editing operations*. 181–190. http://www.statmt.org/wmt12/

Koponen, M. (2015). How to teach machine translation post-editing ? Experiences from a post-editing course. *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)*.

Koponen, M. (2018). Learning to post-edit: An analysis of post-editing quality and processes of translation students. *International Association for Translation and Intercultural Studies (IATIS) 6th International Conference*, *July*. https://doi.org/10.13140/RG.2.2.24675.04648

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. *AMTA Workshop on Postediting Technology and Practice*, *47*(3), 11–20. https://doi.org/10.1111/j.1469-8986.2009.00947.x.Pupillometry

Koponen, M., & Salmi, L. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 137–148.

Krings, H. P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes* (G. S. Koby, Ed.). The Kent State University Press. https://books.google.com.tr/books?id=vsdPsIXCiWAC

Lacruz, I., & Jääskeläinen, R. (2018). *Innovation and Expansion in Translation Process Research*. John Benjamins Publishing Company.

Lacruz, I., Shreve, G. M., & Angelone, E. (2012). Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, 21–30.

Lardilleux, A., & Lepage, Y. (2017). CharCut: Human-Targeted Character-Based MT Evaluation with Loose Differences. *Proceedings of IWSLT 2017*.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M., & Volk, M. (2013, September). Assessing post-editing efficiency in a realistic translation environment. *Proceedings of the 2nd Workshop on Post-Editing Technology and Practice*. https://aclanthology.org/2013.mtsummit-wptp.10

Lauscher, S. (2000). Translation Quality Assessment. *The Translator*, *6*, 149–168. https://doi.org/10.1080/13556509.2000.10799063

Le, Q. v., & Schuster, M. (2016, September 27). *A Neural Network for Machine Translation, at Production Scale*. Google AI Blog. https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Lilly, P. (2016, September 28). *Google Translate Has Reached Human-Like Accuracy Thanks To Neural Machine Translation Engine*. HotHardware. https://hothardware.com/news/google-translate-human-like-accuracy-neural-machine-translation-engine

Lommel, A. (2016, June 22). *MT is Changing the Industry, Just Not in the Way Mainstream Media Thinks It Will*. https://csa-research.com/Blogs-Events/Blog/MT-is-Changing-the-Industry-Just-Not-in-the-Way-Mainstream-Media-Thinks-It-Will

Massardo, I., van der Meer, J., O'Brien, S., Hollowood, F., Aranberri, N., & Drescher, K. (2016). *MT POST-EDITING GUIDELINES*. TAUS Signature Editions.

Milengo GmbH. (2019, January 30). *How NMT-Based Translation Services Can Reduce Enterprise Translation Costs by up to 80%*. Slator. https://slator.com/how-nmt-based-translation-services-can-reduce-enterprise-translation-costs-by-up-to-80/

Moorkens, J., O'Brien, S., da Silva, I. A. L., de Lima Fonseca, N. B., & Alves, F. (2015). Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, *29*(3–4), 267–284. https://doi.org/10.1007/s10590-015-9175-2

Nitzke, J., & Oster, K. (2016). Comparing Translation and Post-editing: An Annotation Schema for Activity Units. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research. New Frontiers in Translation Studies.* Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-20358-4_14

O'Brien, S. (2002). Teaching Post-Editing: A Proposal for Course Content. *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*. https://aclanthology.org/2002.eamt-1.11

O'Brien, S. (2006). Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. *Across Languages and Cultures*, *7*(1), 1–21. https://doi.org/10.1556/acr.7.2006.1.1

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces*, *1*, 101–122. https://doi.org/10.1075/ts.1.05obr

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Plitt, M., & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, *93*(1), 7–16. https://doi.org/10.2478/v10108-010-0010-x

Poibeau, T. (2017). *Machine Translation*. Cambridge, MA: The MIT Press.

Popović, M., Lommel, A., Burchardt, A., Avramidis, E., & Uszkoreit, H. (2014). Relations between different types of post-editing operations, cognitive effort and temporal effort. *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 191–198. https://www.researchgate.net/publication/268446855%0Ahttp://www.dfki.de/web/forschung/publikationen/renameFileForDownload?filename=finalVersion48.pdf&file_id=uploads_2255

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta: Journal Des Traducteurs*, *58*, 487. https://doi.org/10.7202/1025047ar

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rico, C., & Torrejón, E. (2012). Skills and Profile of the New Role of the Translator as MT Post-editor. *Revista Tradumàtica: Tecnologies de La Traducció*, *2012*(10), 166–178. http://revistes.uab.cat/http://revistes.uab.cat/tradumatica

Sánchez-Gijón, P., Moorkens, J., & Way, A. (2019). Post-editing neural machine translation versus translation memory segments. *Machine Translation*, *33*(1–2), 31–59. https://doi.org/10.1007/s10590-019-09232-x

Sankaravelayuthan, R., & Vasuki, G. (2013). *English to Tamil machine translation system using parallel corpus*. LAP LAMBERT Academic Publishing. https://www.perlego.com/book/3413858/english-to-tamil-machine-translation-system-using-parallel-corpus-pdf

Schwartz, L. (2018). The history and promise of machine translation. *American Translators Association Scholarly Monograph Series*, *18*, 161–190. https://doi.org/10.1075/ata.18.08sch

Screen, B. (2019). What effect does post-editing have on the translation product from an end- user's perspective? *The Journal of Specialised Translation*, *500*(31), 133–157.

Sin-wai, C. (2017). *The Future of Translation Technology: Towards a World without Babel*. Routledge.

Stefaniak, K. (n.d.). *Post-editing in DGT*. European Commission, Directorate-General for Translation.

Sukkhwan, A., & Sripetpun, W. (2014). *Students' attitudes and behaviors towards the use of google translate* [Unpublished master's thesis]. Prince of Songkla University.

Teixeira, C. S. C. (2014). Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, 45–59. https://aclanthology.org/2014.amta-wptp.4

Temizöz, Ö. (2012). Machine translation and postediting. *Est-Translationstudies.Org*, 19. http://www.est-translationstudies.org/intranet/research/MT.pdf

Temizöz, Ö. (2016). Postediting machine translation output: subject-matter experts versus professional translators. *Perspectives: Studies in Translatology*, *24*(4), 646–665. https://doi.org/10.1080/0907676X.2015.1119862

Thames, J. (2019, June 24). *Machine Translation*. LanguageSolutions. https://langsolinc.com/machine-translation/

Toral, A., Wieling, M., & Way, A. (2018). Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities*, *5*. https://doi.org/10.3389/fdigh.2018.00009

*Trados Studio - Translation Software*. (n.d.). Retrieved May 18, 2022, from https://www.trados.com/products/trados-studio/

Turovsky, B. (2016, November 15). *Found in translation: More accurate, fluent sentences in Google Translate*. The Keyword. https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

Vanroy, B., de Clercq, O., & Macken, L. (2019). Correlating process and product data to get an insight into translation difficulty. *Perspectives: Studies in Translation Theory and Practice*, *27*(6), 924–941. https://doi.org/10.1080/0907676X.2019.1594319

Vieira, L. N. (2016). How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation*, *30*(1–2), 41–62. https://doi.org/10.1007/s10590-016-9188-5

Vieira, L. N. (2018). Automation anxiety and translators. *Translation Studies*, *0*(0), 1–21. https://doi.org/10.1080/14781700.2018.1543613

Yamada, M. (2019). The impact of Google Neural Machine Translation on Post-editing by student translators. *The Journal of Specialised Translation*, *31*, 87–106.

# APPENDIX A
# PRE-TEST QUESTIONNAIRE

## ANKET

Lütfen formun üzerine sizi tanımlayıcı bilgiler (ad, soyad gibi) yazmayın. **Aksi belirtilmedikçe** yalnızca 1 (bir) seçenek işaretleyin.

1. Yaşınız: _____________

2. Eğitim düzeyiniz (lütfen işaretleyin): □ Lisans □ Yüksek Lisans □ Doktora

3. Profesyonel çeviri deneyiminiz: □ 0 yıl □ 1-3 yıl □ 3-5 yıl □ 5-10 yıl

4. Çeviri yaptığınız dil çiftleri: ___________________________________

5. Çeviri aracı kullanıyor musunuz? □ Evet □ Hayır

6. Evet ise düzenli olarak kullandığınız çeviri araçlarını işaretleyin. (Birden fazla işaretleyebilirsiniz.)

□ SDL Trados
□ MemoQ
□ Memsource
□ Smartcat
□ Diğer: ___________________________________

7. Çeviri projelerinizde makine çevirisi araçlarından faydalanıyor musunuz? (Ör. Google Translate)

□ Evet □ Hayır

8. Evet ise makine çevirisini ne sıklıkla kullanıyorsunuz?
□ Her zaman
□ Neredeyse her zaman
□ Ara sıra
□ Nadiren
□ Asla

9. Makine çevirisini ne amaçla kullanıyorsunuz? (Birden fazla işaretleyebilirsiniz.)
□ Sözlük
□ Çevirdiğim metnin bağlamını anlama
□ Makine çevirisi üzerinde postediting yapma (“çeviri” projelerinde)

10. Makine çevirisi hakkında görüşünüze en çok uyan seçeneği işaretleyin.

☐ Gelecekte her türlü çevirinin makine çevirisiyle yapılacağına inanıyorum.

☐ Makine çevirisi faydalı bir araç ama insan çevirmenlerin müdahalesi her zaman gerekli.

☐ Makine çevirisinin henüz yeterince gelişmediğini düşünüyorum.

☐ Makine çevirisinin kullanılmaması gerektiğine inanıyorum.

11. Bence makine çevirisi…

☐ Yeterli kapasiteye sahip değil.

☐ Her projede kullanılabilir.

☐ İnsan çevirmenler düzelttiği sürece kullanışlıdır.

☐ Yanlış çeviri riski nedeniyle kullanılmamalıdır.

12. Daha önce postediting projelerinde yer aldınız mı?
☐ Evet ☐ Hayır

13. Evet ise tamamladığınız proje sayısını belirtin: _________________________

Anketi tamamladığınız için teşekkür ederiz. Lütfen sonraki adımlar için araştırmacının talimatlarını izleyin.

# APPENDIX B

# POST-TEST QUESTIONNAIRE

## ANKET

Lütfen formun üzerine sizi tanımlayıcı bilgiler (ad, soyad gibi) yazmayın. **Aksi belirtilmedikçe** yalnızca 1 (bir) seçenek işaretleyin.

Bu anket tamamladığınız alıştırmada harcadığınız eforu ölçmek üzere tasarlanmıştır. Lütfen verilen seçeneklerden size en uygun olanı işaretleyin.

1. Düzenlediğim segmentlerde çevirilerin kaynağını (makine veya insan) anladığımı düşünüyorum.

☐ Evet ☐ Hayır ☐ Emin değilim

2. Düzenlediğim segmentlerde çeviri kalitesi…

☐ Çok iyiydi

☐ İyiydi

☐ İdare ederdi

☐ Kötüydü

☐ Çok kötüydü

Aşağıda 1-9 arasında (**1 = "çok, çok düşük"; 9 = "çok, çok yüksek"**) cevaplamanızı istediğimiz sorular yer almaktadır. Lütfen cümleleri dikkatlice okuyup ölçekte size en uygun dereceyi altındaki alanda işaretleyin.

3. Bu dosya üzerinde çalışırken harcadığım zihinsel efor…

| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

4. Düzeltme yaparken harcadığım zihinsel efor…

| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

5. Düzeltme yaptığım segmentlere kıyasla boş cümleleri çevirirken harcadığım zihinsel efor…

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |

6. Dosyanın kolaylık düzeyi…

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |

7. Düzeltme yaptığım cümlelerin kolaylık düzeyi…

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |

8. Sıfırdan çevirdiğim cümlelerin kolaylık düzeyi…

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |

9. Alıştırma hakkında size uygun olan ifadeyi seçin.

□ Düzeltme yaptığım cümleleri baştan çevirsem daha az efor harcardım.

□ Halihazırda çevrilmiş cümlelerin bulunması bana hız kazandırdı.

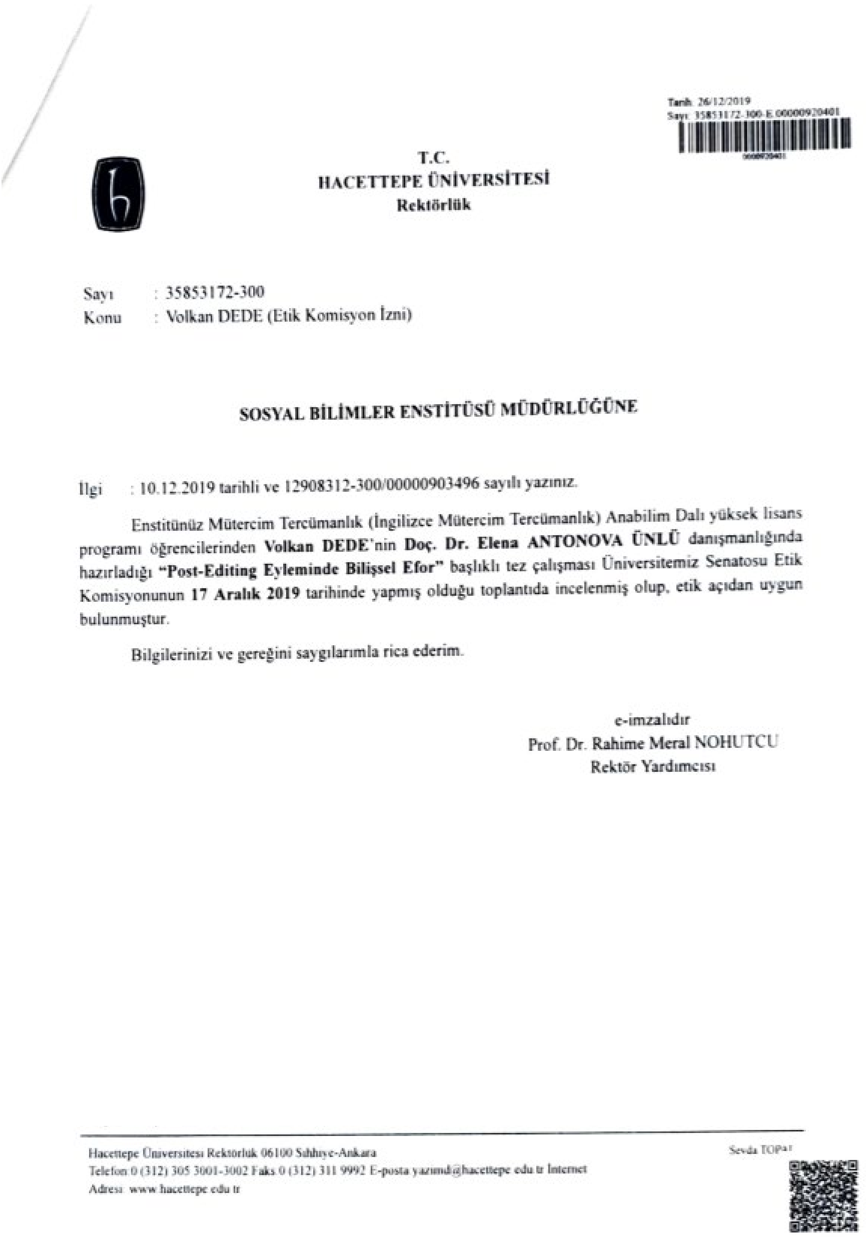
10. Aşağıdaki ifadelerden hangisi sizin için doğru?

□ Düzenlediğim cümlelerin kalitesi işimi zorlaştırdı.

□ Çevrilmiş cümleler olmasa çeviriyi daha uzun sürede bitirirdim.

# APPENDIX C

# ETHICS BOARD APPROVAL FORM



Tarih: 26/12/2019
Sayı: 35853172-300-E.00000920401

T.C.
HACETTEPE ÜNİVERSİTESİ
Rektörlük

Sayı : 35853172-300
Konu : Volkan DEDE (Etik Komisyon İzni)

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 10.12.2019 tarihli ve 12908312-300/00000903496 sayılı yazınız.

Enstitünüz Mütercim Tercümanlık (İngilizce Mütercim Tercümanlık) Anabilim Dalı yüksek lisans programı öğrencilerinden **Volkan DEDE**'nin **Doç. Dr. Elena ANTONOVA ÜNLÜ** danışmanlığında hazırladığı **"Post-Editing Eyleminde Bilişsel Efor"** başlıklı tez çalışması Üniversitemiz Senatosu Etik Komisyonunun **17 Aralık 2019** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

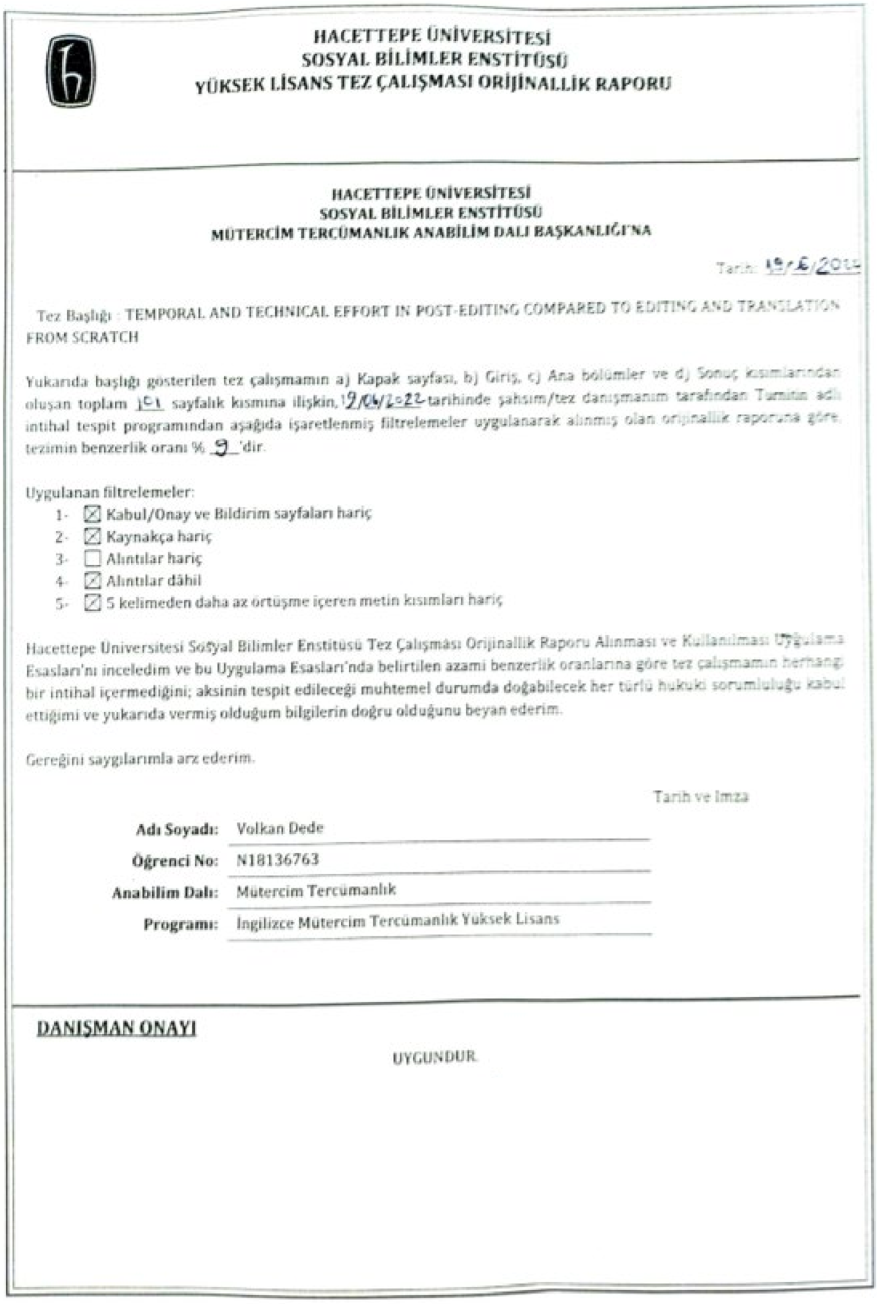
Bilgilerinizi ve gereğini saygılarımla rica ederim.

e-imzalıdır
Prof. Dr. Rahime Meral NOHUTCU
Rektör Yardımcısı

Hacettepe Üniversitesi Rektörlük 06100 Sıhhiye-Ankara
Telefon 0 (312) 305 3001-3002 Faks 0 (312) 311 9992 E-posta yazımd@hacettepe edu tr İnternet
Adresi: www.hacettepe edu tr

Sevda TOPAL

# APPENDIX D
# ORIGINALITY REPORT



**HACETTEPE ÜNİVERSİTESİ**
**SOSYAL BİLİMLER ENSTİTÜSÜ**
**YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU**

**HACETTEPE ÜNİVERSİTESİ**
**SOSYAL BİLİMLER ENSTİTÜSÜ**
**MÜTERCİM TERCÜMANLIK ANABİLİM DALI BAŞKANLIĞI'NA**

Tarih: 19/6/2022

Tez Başlığı : TEMPORAL AND TECHNICAL EFFORT IN POST-EDITING COMPARED TO EDITING AND TRANSLATION FROM SCRATCH

Yukarıda başlığı gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler ve d) Sonuç kısımlarından oluşan toplam 101 sayfalık kısmına ilişkin, 19/04/2022 tarihinde şahsım/tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda işaretlenmiş filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 9 'dir.

Uygulanan filtrelemeler:

1- ☒ Kabul/Onay ve Bildirim sayfaları hariç
2- ☒ Kaynakça hariç
3- ☐ Alıntılar hariç
4- ☒ Alıntılar dâhil
5- ☒ 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

**Adı Soyadı:** Volkan Dede

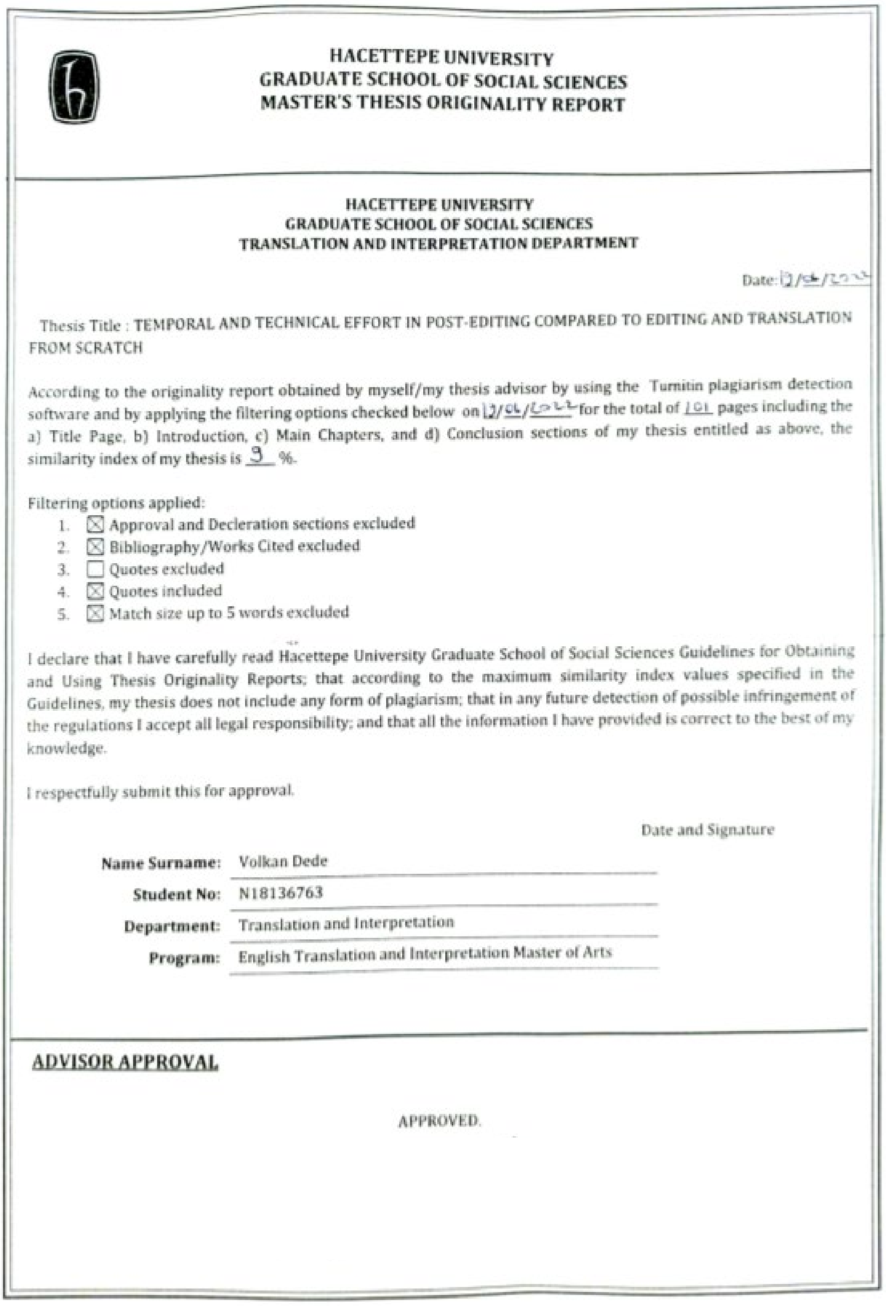
**Öğrenci No:** N18136763

**Anabilim Dalı:** Mütercim Tercümanlık

**Programı:** İngilizce Mütercim Tercümanlık Yüksek Lisans

**DANIŞMAN ONAYI**

UYGUNDUR.



HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
MASTER'S THESIS ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
TRANSLATION AND INTERPRETATION DEPARTMENT

Date: 12/06/2022

Thesis Title : TEMPORAL AND TECHNICAL EFFORT IN POST-EDITING COMPARED TO EDITING AND TRANSLATION FROM SCRATCH

According to the originality report obtained by myself/my thesis advisor by using the Turnitin plagiarism detection software and by applying the filtering options checked below on 12/06/2022 for the total of 101 pages including the a) Title Page, b) Introduction, c) Main Chapters, and d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 9 %.

Filtering options applied:

1. ☒ Approval and Decleration sections excluded
2. ☒ Bibliography/Works Cited excluded
3. □ Quotes excluded
4. ☒ Quotes included
5. ☒ Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Social Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Date and Signature

**Name Surname:** Volkan Dede

**Student No:** N18136763

**Department:** Translation and Interpretation

**Program:** English Translation and Interpretation Master of Arts

**ADVISOR APPROVAL**

APPROVED.