The Effect of Error Type on Pause Length in Post-Editing Machine Translation Output

Adrian Probst
ANR 456116

Master's Thesis
Communication and Information Sciences
Specialization Human Aspects of Information Technology (HAIT)

School of Humanities
Tilburg University, Tilburg

Supervisor: Dr. M.M. van Zaanen
Second Reader: Dr. A.T. Hendrickson
July 2017

Table of Contents

DISCLAIMER:

This publication has emanated from research supported by the TraMOOC project (Translation for Massive Open Online Courses), funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under Grant Agreement Number 64433.

Abstract

Post-editing machine translated output is a common practice in the translation industry. Post-editing refers to the editing of machine translated text to publishable content. Especially for large amounts of text it can save various resources. However, not every type of error in post-editing is treated the same way and takes the same amount of time to edit. Therefore, it is hard to estimate the editing time, as it largely depends on the error types that the software produces. Not only the post-editing takes more time for certain errors, but also the pause time can vary. Pause time represents the time after the last cursor positioning and before the first editing movement. Currently, the two most commonly used approaches in machine translation are Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). This thesis focuses on three key constructs: (1) the amount of errors and the error types produced by SMT and NMT, (2) the pause length prior to editing errors in SMT and NMT as well as (3) the pause length prior to five annotated error types. In a first analysis, it is shown that NMT software produces output with fewer absolute errors than SMT software. Moreover, in a second analysis, it is shown that, in general, the pause length in post-editing prior to errors is shorter for NMT output than for SMT output. In addition, an exploratory approach is taken in order to investigate whether the five annotated error types yield differences in the pause length prior to editing. The results of this analysis show that *Addition* errors and *Inflectional Morphology* errors require the shortest pause length prior to editing the error, while a *Mistranslation* error evokes the longest pause prior to the edit.

1. Introduction

Since translations are becoming ever so important in today's globalized and interconnected world and the quality of high-end machine translation (MT) software has improved remarkably over the last decade, it is no surprise that making use of machine translated output has become a regular practice in the translation industry.

1.1 Applications of MT Post-Editing

Machine translated output on its own cannot – and will not in the foreseeable future – reach the required standard for it to be readily publishable text. Post-editing using professional translators is the bridge to close this gap. Post-editing refers to correcting machine translated output in such a way that the text becomes of publishable quality.

There are two main approaches of professional use of MT software where post-editing comes into play. One approach that has become more and more popular in recent years is that language providers (e.g. translation agencies, supply chain operators, etc.) purchase or develop an MT software package and pre-translate all their texts prior to offering them to language professionals for post-editing. This is time and cost efficient for both the language provider as well as the end client. However, there have been controversial opinions on whether the saving time and cost aspect is truly valuable. Garcia (2011), in fact, found that productivity gains with post-editing are nearly insignificant, but the quality of the output in this study was notably better than manually translated output.

A second approach of professional use of MT with post-editing is the use of Computer-Aided Translation (CAT) tools with incorporated MT engines. CAT tools have become a must for every professional translator. They are computer programs that support the translators in their work by saving certain expressions into translation memories and re-using them when the same word comes up. This not only saves time for the translator, but it also reduces costs for the end client and it guarantees coherence. Many established CAT tools now have integrated MT engines that make it possible to pre-translate text and post-edit the outcome to one's perceived perfection. Examples of such tools with integrated MT options are SDL Trados Studio (http://www.sdl.com/), memoQ (https://www.memoq.com/en) or Across (http://www.across.net/en/), among others.

1.2 Cognitive Load of Error Types

The process of post-editing consists of three major efforts: temporal, cognitive, and technical efforts (Krings & Koby, 2001). Temporal effort describes the time that it takes to transform machine translated text into publishable text. Cognitive effort contains the cognitive processes it takes to correct certain errors. Cognitive effort is hard to observe and measure, as it is unsure when and how it exactly takes place. Technical effort represents the actual correction that takes place (addition, deletion, etc.). As can be seen, it is difficult to make a concise distinction between these concepts since the technical and cognitive efforts are taking place within the temporal effort. This thesis is mainly interested in the cognitive effort and in one aspect in particular: the pauses prior to editing that might be caused by cognitive effort among other things (O'Brien, 2006). However, not all error types cause the same amount of cognitive effort. Koponen (2012) suggests that it is, in fact, possible that certain errors that require little technical effort may cause a high cognitive effort and vice versa.

The error annotations from the data that this thesis is built upon are restricted to the five categories of *Inflectional Morphology, Word Order, Omission, Addition,* and *Mistranslation* (explanations and examples to each error will be given in the next chapter). This is because the annotation tool (PET), that was used to annotate the errors, can only distinguish between these five errors.

Previous research attempted to categorize error types in certain groups and to assign them a level of perceived effort that it takes to correct them. Temnikova (2010), for example, introduced a cognitive ranking in which she suggests that errors on the morphological level are easiest to correct, followed by errors on the lexical level and the syntactical level. Errors on the morphological level refer to errors that occur within a word that belongs in the sentence, but not in this form (e.g. incorrect plural). They can usually be fixed rather quickly, as they have no implications on the sentence as a whole. Errors on the lexical level refer to errors that only concern one word that is wrong within a sentence (e.g. incorrect word, missing word, etc.). Syntactic level errors refer to errors that have an effect on the whole sentence (e.g. punctuation, word order, etc.).

There is no consensus in the literature on which error types require how much cognitive effort. The same goes for a concrete categorization of error types. Since this thesis is restricted to five error types, there will not be a predefined categorization of cognitive effort. Therefore, the current thesis pursues an exploratory approach towards the five different error types. This approach aims to discover differences in pause lengths for the different error types prior to editing the errors rather than categorizing them and formulating hypotheses.

1.3 Pauses in Post-Editing

Post-editing has been subject to numerous previous research (Allen, 2003; Arenas, 2008; Guerberof, 2009; Krings & Koby, 2001; Löffler-Laurian, 1986; McElhaney & Vasconcellos, 1988; Melby, Housley, Fields & Tuioti, 2012; Plitt & Masselot, 2010). It can be summarized that most studies focus on what the translators are doing in post-editing. However, only few studies focus on the pauses in post-editing, when the translator is thinking prior to starting the editing process (Krings & Koby, 2001; Lacruz, Shreve & Angelone, 2012; O’Brien, 2006). O’Brien (2006) investigated whether the pause times differ for sentences that have a higher machine translatability compared to sentences that seem to be less suitable for MT. Lacruz et al. (2012) based their research on O’Brien’s findings and introduced a new, more representative measurement called “average pause ratio”. More on their findings can be found in the next chapter.

Interestingly, a combination of both error types and pause time has, to the best of our knowledge, only been investigated to some extent by Popovic et al. (2014). In their study, Popovic et al. found that lexical choice and word order errors occurred most frequently in the MT output that was tested. Moreover, they found that lexical edits took the most time, whereas word order errors took the least amount of time to edit.

This thesis attempts to extend the body of knowledge in this intertwined field of research, combining error types and pause time in post-editing. Based on time code information of post-edited machine translated content, the pauses prior to each error will be assessed and examined in order to make a statement about whether there is an effect on the pause length depending on the type of error it precedes.

1.4 MOOC Domain

The data used in this thesis is part of a project funded by the European Commission that aims to provide MT systems that translate Massive Open Online Courses (MOOCs) by using a state-of-the-art Neural Machine Translation (NMT) software. This project, called Translation for Massive Open Online Courses (TraMOOC), is a Horizon 2020 collaboration led by Humboldt Universität zu Berlin. The team consists of ten partners from six different European countries –

one of which being Tilburg University. The main goal of this project is to provide high-quality machine translated content of openly available text in the educational domain (e.g. subtitles of lectures, forum posts, etc.).

This data was collected as part of a study conducted by Dublin City University, which attempted to compare output from SMT to NMT. TraMOOC recently shifted their translation approach from SMT to NMT.

1.5 Research Questions and Hypotheses

The previous sections introduced the main constructs that this thesis will focus on: the amount of absolute errors and error types in SMT and NMT tools as well as the pause lengths when post-editing prior to these errors and error types. Therefore, this thesis attempts to answer the following research questions:

(RQ1) Do SMT and NMT tools differ in the amount of errors that they produce in the machine translated output?

(RQ2) Do SMT and NMT tools differ in the pause length prior to editing errors?

(RQ3) What is the effect of certain error types on the respective length of the pause that immediately precedes the errors?

In order to answer these research questions, three hypotheses were set up. Hypotheses 1 attempts to answer Research Question 1 (H_1 → RQ_1), Hypothesis 2 refers to Research Question 2 (H_2 → RQ_2) and Hypothesis 3 attempts to answer Research Question 3 (H_3 → RQ_3). The hypotheses are largely based on theories and concepts introduced in chapter 2 of this thesis:

(H_1) NMT generated output is expected to contain fewer absolute errors than SMT generated output.

(H2) The pause length prior to errors is expected to be larger in SMT translated output compared to NMT translated output.

(H3) There is a difference in pause length prior to post-editing across certain error types (exploratory approach).

1.6 Thesis Structure

The structure of this thesis is organized in six chapters. Chapter 2 provides background information for the hypotheses by outlining previous research in the fields of MT, post-editing, error types in post-editing and pauses in post-editing. Chapter 3 describes the methods that are used in this thesis to answer the research questions. The analysis of the error types provides an indication for anyone who is interested in purchasing an MT engine, but is unsure whether to invest in a SMT or NMT tool. In addition, this is useful for the TraMOOC project as well, since the platform has recently been shifted from a SMT to a NMT platform. First of all, this thesis will outline which platform delivers fewer errors in MT output. Furthermore, it will be investigated whether the platform with fewer errors also requires less pause time in post-editing. The analysis of the pause length prior to error types will provide an interesting insight for developers of such tools. Ultimately, MT is always about trying to save resources. Chapter 4 describes the result of the analysis and answers the hypotheses, and chapter 5 reflects on these results and attempts to explain the findings. Eventually, chapter 6 summarizes the study, outlines its limitations and suggests further research that is needed in this domain. Both specific research topics as well as general areas for further investigation are presented.

2. Theoretical Background

This chapter outlines and reviews previous work that has been conducted in the fields of MT, post-editing in MT, error types in post-editing as well as pauses in post-editing. Furthermore, different approaches to MT and its evaluation are explored, as well as different opinions with regards to frequent error types in certain target languages. Firstly, however, the domain, which this research can be placed in, is briefly explained in the following section.

2.1 MOOC and TraMOOC

This thesis contributes to the TraMOOC project (Translation for Massive Open Online Courses), which is funded by the European Commission. TraMOOC is a Horizon 2020 collaborative project led by Humboldt Universität zu Berlin that brings together highly skilled researchers and developers from all over Europe to work together in order to achieve its main goal: the development of a platform that provides high quality MTs from English into eleven languages (DE, IT, PT, EL, NL, CS, BG, HR, PL, RU, ZH) of educational texts, the so-called MOOCs. MOOCs are educational texts that are readily available online. The MOOC resources are an ever-growing body of educational texts from subtitles of lectures to forum posts, among others. The main problem is that, so far, the vast majority of MOOC content is only available in English. Some MOOC texts are rather straightforward to translate using MT engines (e.g. lecture slides, subtitles, etc.) as they are composed in correctly written English. However, there are many texts in the MOOC domain that pose great difficulties for MT tools (e.g. forum posts, blogs, comments, etc.). This is due to the fact that the language that is used is mostly spoken and includes abbreviation, slang as well as elliptical structures. Another difficulty in the MOOC domain is that some course material uses extremely domain-specific terminology. This means that terms are being used that occur only in highly specific situations, which converts into a low term frequency. This makes it harder for any SMT or NMT tool to translate these terms. TraMOOC took on the task to make this content accessible to the non-English speaking world.

2.2 Machine Translation

MT refers to the field that uses software to translate a text from a source language into a target language. The first ideas to use computers for translations trace back to Andrew Booth and Warren Weaver in the mid-1940s (Hutchins & Lovtsky, 2000). In 1951, Yehoshua Bar-Hillel from the Massachusetts Institute of Technology was appointed the first researcher in the field of MT (Ramlow, 2009). It was already clear at this point, that the development of tools for publishable MT output will be close to impossible and that the need for human intervention by post-editing will be indispensable (Hutchins, 2007).

There have been several approaches to MT over the last 60 years, with the latest – and currently most popular ones – being Statistical Machine Translation and Neural Machine Translation. The TraMOOC project initially started out with building a SMT-based system. In 2016, TraMOOC switched from SMT to NMT. This thesis also focuses on the differences between these two approaches. Therefore, they will be briefly explained in the following subsections. It is worth noting, however, that these are not the only two approaches. Various other systems have been explored and used over the years including Rule-Based Machine Translation and Phrase-Based Machine Translation (Chiang, 2005; Eisele, et al., 2008; Kamran, 2013).

2.1.1 Statistical Machine Translation (SMT)

SMT systems work on probability calculations. The developer of such a system creates a model which is able to learn rules and patterns based on bilingual text corpora. This model then predicts the words that need to be translated. A large benefit of SMT systems are the fact that developers are able to improve the quality of the system by slightly adapting the underlying model (Al-Onaizan et al., 1999). This is a cost and time efficient way of improving accuracy. Another advantage is that SMT accuracy is typically increased when more data is available to learn from. The models behind SMT systems rely on parallel text corpora in order to find word alignments. Word alignments refer to the mapping of source words and target words in sentences. The models then compute the frequency with which words co-occur. Every word that the model registers from the training data is a potential translation (Jurafsky & Martin, 2014). This means that SMT requires fewer linguistic features as it focuses on statistical modelling.

2.1.2 Neural Machine Translation (NMT)

NMT systems, which contain neural networks that can be trained specifically for translating, is the latest development in the field of MT and has picked up increasing interests (Bahdanau, Cho & Bengio, 2015; Sutskever, Vinyals & Le, 2014). Since Google announced their new MT system GNMT (Google’s Neural Machine Translation system) in September 2016, NMT is considered the state-of-the-art in the field (Wu et al., 2016).

Deselaers et al. (2009) were the pioneers in applying Deep Learning approaches to MT for an Arabic-English translation. Their approach was based on Deep Belief Networks (DBN), models that can be trained on multiple layers. Each layer builds on the output of the previous layer until the architecture of the model is deep enough.

Nowadays, NMT systems usually consist of two recurrent neural networks, one for the input source text and one for the generated output target text (Wu et al., 2016). NMT systems contain an encoder as well as a decoder. The encoder’s job is to represent the input in a way that the decoder is able to translate text according to this representation. This enables the system to directly learn the mapping of an input sentence to its translated output sentence.

One advantage of NMT is especially the fact that it requires considerably less memory space than SMT tools (Cho, Van Merriënboer, Bahdanau & Bengio, 2014). Moreover, NMT is able to learn the difficulties that come with – for instance – the MOOCs that were mentioned earlier, such as abbreviations, slang, spoken language.

2.3 Post-Editing in Machine Translation

Post-editing MT output has been subject to numerous studies. Some studies focused on the different types of post-editing (Allen, 2003; Löffler-Laurian, 1986; McElhaney & Vasconcellos, 1988), others focused largely on post-editing efforts or on the quality of the post-edited content (Arenas, 2008; Guerberof, 2009; Melby, Housley, Fields & Tuioti, 2012; Plitt & Masselot, 2010). Melby et al. (2012) investigated the use of "formalized structured translation specifications (FSTS)" in order to assess the quality of post-editors. The aim of this research was to test whether potential evaluators are able to assess and rate the quality of post-edited text. Using two software programs developed for this research, seventeen non-experts were asked to assess the quality of each text. Melby et al. found that the evaluators were able to perform the quality assessment based on the FSTS specifications.

Guerberof (2009) looked into the integrated MT systems in CAT tools used by professional translators. In order to assess the effect on quality, Guerberof conducted an experiment using eight professional translators. The participants were asked to translate 265 words from scratch, 264 words from the translation-memory 80% - 90% fuzzy-matches (i.e. translations that are not 100% complete from the translation memory, but based on their context, the CAT tool suggests a translation) and post-edit 262 words that were SMT translated output. The translators were not aware of these categorizations. Later on, these texts were revised and the errors were counted. Guerberof found that more than 50% of the total errors occurred in words in the segments coming from fuzzy-matches. 27% of the total errors were made in MT output and 21% in newly translated segments. Guerberof explains these results by stating that text coming from translation memories would flow better and the translators therefore would not consult the source text that much to spot mistakes, whereas the MT output would not be very fluent which made the translators check the sentences more carefully.

2.4 Error Types in Post-Editing

The categorization of error types is usually a part of evaluating the quality of an MT tool. However, not every error type requires the same amount of editing effort. Koponen (2012) conducted a study investigating technical and cognitive effort in post-editing by looking at cases where the automatic evaluation differs from human evaluation. Koponen suggests that shorter editing times are linked with errors that require fewer cognitive effort such as morphology errors, incorrect words and synonym substitutions, whereas longer editing times are associated with more cognitive effort such as incorrect part-of-speech or word order errors.

Temnikova (2010) investigated the cognitive efforts that translators face while post-editing certain error types. Temnikova based her error categories on the error classification developed by Vilar et al. (2006). In addition to this error classification, Temnikova developed a MT error ranking system that ranks certain error types from one (being the easiest to correct) to ten (being the hardest to correct).

As previously mentioned, there is not a general agreement on the categorization of error types to be found in literature. Therefore, the current thesis pursues an exploratory approach towards the five different error types. This approach aims at discovering differences in pause lengths prior to editing the error types rather than categorizing them and formulating hypotheses. The five error types found in the data are *Inflectional Morphology*, *Word Order*, *Addition*, *Omission and Mistranslation*. Table 1 explains each error type from the dataset and provides an example for each.

Table 1

Explanation of Error Types Including Examples.

Error Type	**Explanation**	**Example Source**	**Example Target**	**Example Target corrected**	**Comment**
Inflectional Morphology	Relates to rules that govern form. Something went wrong with the forming of words in certain grammatical categories (e.g. plural form).	What **is** the most essential information in Business Analyst CV?	Was **ist** die wichtigsten Informationen in Business Analyst Lebenslauf?	Was **sind** die wichtigsten Informationen im Business Analyst Lebenslauf?	In this case, the word "be" in the German output was conjugated in the singular form "ist" instead of the plural form "sind".
Word order	Relates to words that are in the wrong order within the sentence.	Another one can be seen **at 1:40**.	Ein anderes gesehen werden kann, **um 1: 40**.	Ein anderes kann **um 1: 40** gesehen werden.	In this case, the time stamp "1: 40" was put at the end of the sentence like in English. In German, this indication needs to come earlier.
Addition	Relates to words that were added to the target even though they are not needed.	SPELLING **mistake** lol	Rechtschreib-fehler **Fehler** lol	Rechtschreib-fehler lol	In this case, a second "Fehler" (mistake) was added.
Omission	Relates to a word in the source that is missing in the target.	What **core** skills I have to learn?	Welche Fähigkeiten muss ich lernen?	Welche **Grund**fähigkeiten muss ich lernen?	In this case, the word "core" was not translated.
Mis-translation	Relates to false translations of the source.	Have I to get any certifications ?	Habe ich irgendwelche Zertifikate bekommen?	Brauche ich irgendwelche Zertifikate?	In this case, the meaning was wrongfully transferred due to bad grammar in the source.

It is assumed, based on previous research, that *Inflectional Morphology* errors tend to be the most frequent in English-German translations and that they take the least effort to edit, as they stay within the morphological level, while the rest requires editing on the lexical level (*Addition, Omission, Mistranslation*) or the syntactic level (*Word Order*) (Temnikova, 2010).

2.5 Pauses in Post-Editing

As stated earlier, most studies on post-editing in MT focused on the editing efforts in post-editing. Only a few studies focus on the pauses in post-editing, when the translator is not typing, but thinking prior to the first act of editing (Krings & Koby, 2001; Lacruz et al., 2012; O'Brien, 2006). O'Brien (2006) investigated whether there was a difference in pause time while editing sentences that have a higher machine translatability compared to sentences with a lower translatability. The aim of this study was to investigate the correlation between the source text's machine translatability and the cognitive efforts in the post-editing process.

Pauses were not the only factor that contributed to cognitive effort. O'Brien used the pause analysis in triangulation with the so-called Choice Network Analysis. This study used nine participants to post-edited an English text that was translated into German with a MT tool. The text contained both sentences that were suited for MT and sentences that were not suited for MT. The log data was recorded using Translog. O'Brien then calculated the so-called "pause ratio" which is the total time spent editing a segment, divided by the total time spent pausing in a segment. The study found no significant difference in "pause ratio" between the two sets of sentences. According to O'Brien, these results can be explained in two ways: pauses alone are not valid indicators of cognitive load or that there is simply no difference between text that is suited for MT and text that is not suited.

Lacruz et al. (2012) based their research on O’Brien’s findings; however, they introduced a new measurement called the “average pause ratio”, which does not only take the pause time into account, but also the number of pauses and their duration, making the measurement more representative. Lacruz et al. recruited a professional translator with multiple years of translation experience, but no previous experience with post-editing MT output. The translator was asked to post-edit a text that had been translated from English into Spanish with a phrase-based MT system. The keystrokes were recorded using Inputlog. Lacruz et al. defined the cognitive effort of sentences based on the completed edit events that were executed. Sentences were segmented and categorized as more cognitively demanding when four or more completed editing events took place, and as less cognitively demanding when two or fewer completed editing events took place. Eight of the thirteen analyzed segments were more cognitively demanding and five were categorized as less cognitively demanding. Lacruz et al. found that the “average pause ratio” for cognitively more demanding segments was lower than for cognitively less demanding segments. This was because more demanding segments contained many short pauses, whereas in less demanding segments most of the time would be spent in “reading comprehension, problem recognition, and solution evaluation” (p. 6), which generally yields in longer pauses.

Krings & Koby (2001) regard pauses especially useful in defining boundaries in the translation process. In this study, they defined pauses as interruptions that last at least one second. When a pause lasts at least one second, it introduces a new coding unit. They explain that this one second rule was chosen arbitrarily, but that it complied well with the study as the interruption in the verbalization flow could easily be identified.

2.6 Pauses in Post-Editing in Relation to Error Types

A combination of the subsections 2.4 and 2.5 provides the theoretical basis for this thesis. To the best of our knowledge, Popovic et al. (2014) is the only research that to some extent combined the constructs of error type and time. They looked at a total of five different types of post-editing operations, namely *correcting word form, correcting word order, adding omission, deleting addition* and *correcting lexical choice*. The goal of this research was to explore the connection between these five operations with cognitive effort and time. Popovic et al. translated a total of 5,779 sentences from French to English as well as from English to Spanish using a SMT system. Two professional human translators for each language pair were assigned to post-edit the output with the instruction of editing as little as necessary. The editing operations were tracked by using the Hjerson automatic tool for error analysis (Popovic, 2011).

The results of this study suggest that lexical choice and word order were the most frequent errors. Another interesting finding was that the post-editing time for the Spanish text was considerably longer than for the English output. This might, however, be due to the translators. Regarding the temporal effects, Popovic et al. (2014) found that lexical edits took the largest amount of time. In the English output, word order errors took the least time to correct. The Spanish post-editor took the least time for deleting additions from the text.

In sum, it can be said that this research is closely related to the research in the thesis. However, the aspect of pause length related to error types is still unexplored. The aim of this thesis is to extend the body of knowledge in this field.

3. Method

In a first step, this chapter focuses on the tools and data that were used during this research. This is followed by the three measures that were investigated during this thesis: error frequency, pause length prior to errors and pause length prior to error types. The last section of this chapter explains how the dataset was prepared in order to be able to investigate the measures. Moreover, the last section also provides a visual representation of the model of this thesis indicating how the hypotheses are attempted to be answered.

3.1 Tools and Data

3.1.1 Dataset

The dataset that is used in this thesis derives from a study conducted at Dublin City University (Castilho et al., 2017). Castilho et al. (2017) processed a set of sentences from different MOOCs with the current version of the TraMOOC tool that is based on NMT as well as with the previously used platform that was based on SMT. The dataset consists of four language pairs (English into German, Portuguese, Greek and Russian). This thesis focuses solely on the language pair English into German. In the next step, a number of professional human translators (two in the case of German) were asked to assess both the SMT and NMT output as well as to post-edit the output into a publishable text using the Post-Editing Tool (PET). Variables that the translators were asked to annotate included: side-by-side ranking, accuracy and fluency rating and error annotation. More about this evaluation is described in section 3.2. There are four data files available for the language pair English-German. Two files with 500 segments each that were annotated by Translator 1 with all the odd numbered segments being translated by SMT and all the even numbered segments being translated by NMT. Plus, two files with 500 segments each that were annotated by Translator 2 with all the even numbered segments translated by

NMT and the odd numbered segments translated by NMT. This results in a complete dataset of 2,000 translations (1,000 by SMT and 1,000 by NMT). The keystrokes as well as the editing and assessing times were tracked by the Post-Editing Tool (PET).

3.1.2 PET (Post-Editing Tool)

The Post-Editing Tool PET by Wilker Aziz and Lucia Specia is a tool that allows researchers to collect explicit and implicit information about post-editing tasks. The tool was developed with Java-6 libraries (Aziz, Castilho & Specia, 2012). A recent version of Java Virtual Machine is all it takes for the program to run. PET is an open source software that is still in development.

The text to be post-edited is displayed in segments with the source language and the target language side-by-side. It can be set whether the translator sees each segment at a time or whether he or she is able to see the preceding and following segments as well. The data for this thesis was post-edited without seeing preceding and following segments (Castilho et al., 2017). PET collects a number of indicators by default and there are additional indicators that can be enabled if necessary. Some of the information that can be tracked and assessed are editing efforts, quality of translation, adequacy and frequency rating, level of language proficiency and – most importantly for the present thesis – timecodes and time spent on editing and assessing.

3.2 Evaluation

The aforementioned data was evaluated by two professional English to German translators using PET over a two-week period. The translators were asked to assess the quality of the output, to perform error annotation as well as to correct the translations and to transform them into publishable quality. The translators were not familiar with post-editing MT output. However, they went through a short training prior to this task and due to their professional experience, their annotations can be trusted. The segments were presented to the translators in chronological order so that the context would be of help during the process. The translators were provided with a manual of the PET tool in order to know how to perform the task. PET allowed the translator to only always see one segment at a time. This led to the translator not being able to peek at the following segment while not having completed the current segment. This was an important aspect to take into consideration for the present thesis, as the errors of the following segments could not be spotted ahead of time. Moreover, it limited the attention of the translator to the current segment, ensuring that the time recorded was indeed spent on this segment.

3.3 Measures

This thesis investigated three measures: the *error frequency*, the *pause length prior to errors* and the *pause length prior to error types.*

3.3.1 Error Frequency

The *error frequency* was measured by assessing how many absolute errors are produced by SMT and NMT. Based on results from previous research, it is assumed that NMT would produce fewer absolute errors than SMT (Castilho et al., 2017). In addition, we investigate which error types occur how many times for each approach and whether there are differences in the amount of errors that are found per segment.

3.3.2 Pause Length Prior to Errors

The *pause length prior to errors* was assessed using the log data provided by PET that indicated the exact processes that post-editors went through while turning the machine translated output into publishable one. The editing process in the log data was annotated with a time stamp. This time stamp functioned as the basis for this measurement.

Figure 1 visually represents the definition of pause length. This is not in line with the way previous investigators defined pause length. This definition stems from the available resources of the dataset. While taking reading time into account, this was the only option to empirically measure a pause that is immediately linked to a certain error. After the editor read the segment and processed the errors within the sentence, he or she starts the editing process. For the purposes of this thesis, pause length is defined as the time after the last cursor positioning (i.e. clicking into the text or bringing cursor into position with arrow keys) and before the first actual editing movement (e.g. addition, deletion, etc.). In Figure 1, this time is visualized with shading.

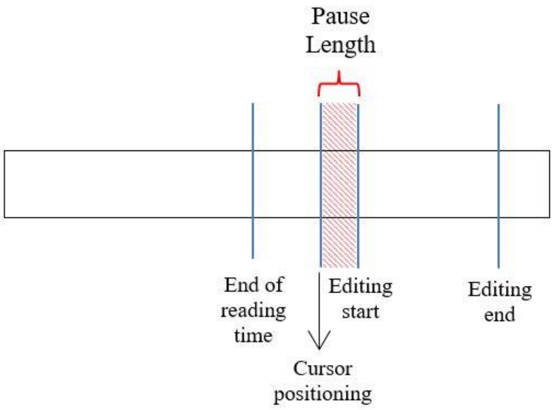

Figure 1. Pause Length Prior to Errors

3.3.3 Pause Length Prior to Error Types

The *pause length prior to error types* was assessed in the same way as the measurement above. However, instead of looking at the errors in general, the pause length was assessed while taking into account the five different annotated error types *Inflectional Morphology, Word Order, Omission, Addition* and *Mistranslation*. It was tested whether there are differences in pause length between the error categories that could potentially be traced back to the cognitive load that certain error types carry.

3.4 Procedure

In a first step, the data files were prepared by removing unnecessary information in order to make it more feasible to conduct the study. The data files contained various bits of information that were not needed for this thesis such as assessment ratings, assessment time logs, correction summaries and various other annotations. The next step was counting the frequency of errors occurring in SMT and NMT output for both Translator 1 and 2 as well as segmenting them into the five different error types based on their annotations. All segments were once pre-translated by the SMT and once by the NMT tool. Each segment was regarded as one individual case in order to be able to conduct statistical tests in a later stage.

Based on the results of the first analysis, a parallel stream of investigation was launched by assessing the pause lengths that occurred in post-editing immediately prior to the corrective action of errors. This investigation was the basis for Hypotheses 2 and 3. After having looked at pause length prior to errors in general, we applied it to the five segmented error types from analysis one in order to see whether there is a significant difference in how long the pauses are in post-editing prior to different error types.

Figure 2 depicts a visual interpretation of the procedure underlying this thesis. It shows that Hypothesis 1, 2 and 3 are parallel investigations; however, Hypothesis 2 and 3 regarding the pause lengths are largely dependent on Hypothesis 1.

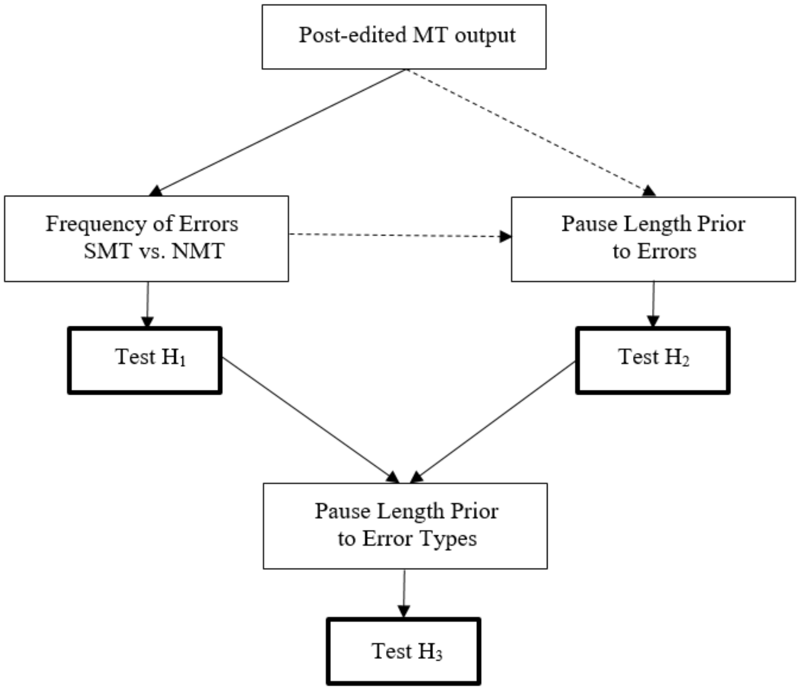
Post-edited MT output
Frequency of Errors SMT vs. NMT
Pause Length Prior to Errors
Test H1
Test H2
Pause Length Prior to Error Types
Test H3

Figure 2. Thesis Procedure Model

4. Results

This chapter presents the results of the investigations that were conducted for this thesis. The first section indicates the errors that were found in the SMT and NMT. Section 4.2 presents whether there are differences in pause lengths in general between SMT and NMT. Lastly, section 4.3 links error types and pause lengths and presents whether there are differences in the pause lengths in relation to the five error types that were annotated in the used data. Based on the literature, we assume that NMT software produces fewer errors than SMT software and that NMT requires smaller pause time prior to editing errors. In addition, it is assumed that *Inflectional Morphology* errors will require the smallest pause length prior to editing, as they should be easiest to correct. Moreover, section 4.3 attempts to predict the results onto the complete dataset to see the total time spent pausing in SMT and NMT. This is an estimation and of exploratory nature and the results should, therefore, be treated with caution.

Even though a direct link between pause length and cognitive effort is not claimed, these results do provide some interesting insight that can be investigated with follow-up studies.

4.1 Error Frequency in SMT and NMT

The four datasets combined are composed of 1,000 different segments and 2,000 translations. The double amount of translations derives from each segment being once translated by SMT and once by NMT. A total of 157 segments needed to be excluded from the data due to misleading factors. Instead of just one set of error annotations, these segments included two or more sets of annotations. After investigating the cause of this difference, it was found that the translators working on these files were able to re-enter a segment after they had already completed it. Since this distorted the initial assessment from the respective segment, it made it impossible to include these into the analyses.

The remaining segments amounted to a total of 916 for SMT and 927 for NMT. The total of errors in the SMT output was 1,536, compared to a total of 1,124 errors in NMT output. The results of this study showed that the NMT approach proportionally yielded a lower number of absolute errors (M = 1.21, SD = 0.03) than the SMT approach (M = 1.68, SD = 0.03).

The assumption of normality was violated both for SMT as well as for NMT as can be seen in the Kolmogorov-Smirnov test D(916) = .225, p < .001 and D(927), p < .001. Due to this violation, the parametric statistical testing using a paired-samples t-test was not possible. For this reason, the non-parametric Mann-Whitney U test was applied in order to detect if this difference was significant. The Mann-Whitney test indicated that the output generated by NMT contained significantly less absolute errors than SMT output (U = 303479.00, p < .001).

For acquiring a better understanding of the whole dataset, Table 2 outlines the numbers for each error type across software approaches. As can be seen, NMT output managed to increase its proportion of *Inflectional Morphology* errors, which generally are easier to fix, as they are errors on the morphological level of text, to 50% compared to the other error types. SMT, on the other hand, seemed to struggle more with word order errors as they make up more than 1/5th of the total errors.

Table 2

Error Frequency Across Error Types and Approaches.

Error Type	**Errors in SMT**	**Errors in NMT**
1. Inflectional Morphology	672 (43.75%)	562 (50.00%)
2. Word Order	352 (22.92%)	171 (15.21%)
3. Omission	115 (7.49%)	73 (6.49%)
4. Addition	35 (2.27%)	29 (2.58%)
5. Mistranslation	362 (23.57%)	289 (25.72%)
Total	1,536 (100.00%)	1,124 (100.00%)

Another factor that had to be taken into account in order to fully comprehend the dataset was the amount of errors that were produced within one segment. NMT produced 742 segments with errors in them, compared to 857 segments for SMT. As Table 3 shows, the newer NMT approach managed to shift most of its segments containing errors to the top two ranks of the list (i.e. one error or two errors). While SMT produced a large number of segments with three errors, NMT kept this number down to a third respectively. As the totals of each approach are not the same in this case, it makes sense to represent the data visually. This way, the error proportion in percentages across the segments is clear. Figure 3 shows the trend that NMT managed to produce more segments of zero or one error, while SMT shows more in the worse categories of two and three errors. Segments with four or five errors are very rare in both cases.

Table 3

Amount of Errors in Segments across SMT and NMT.

Errors in Segment	**Segments in SMT**	**Segments in NMT**
0 errors	59 (6.44%)	185 (19.96%)
1 error	344 (37.55%)	422 (45.52%)
2 errors	358 (39.08%)	265 (28.58%)
3 errors	145 (15.84%)	49 (5.29%)
4 errors	9 (0.98%)	5 (0.54%)
5 errors	1 (0.11%)	1 (0.11%)
Total	916 (100.00%)	927 (100.00%)

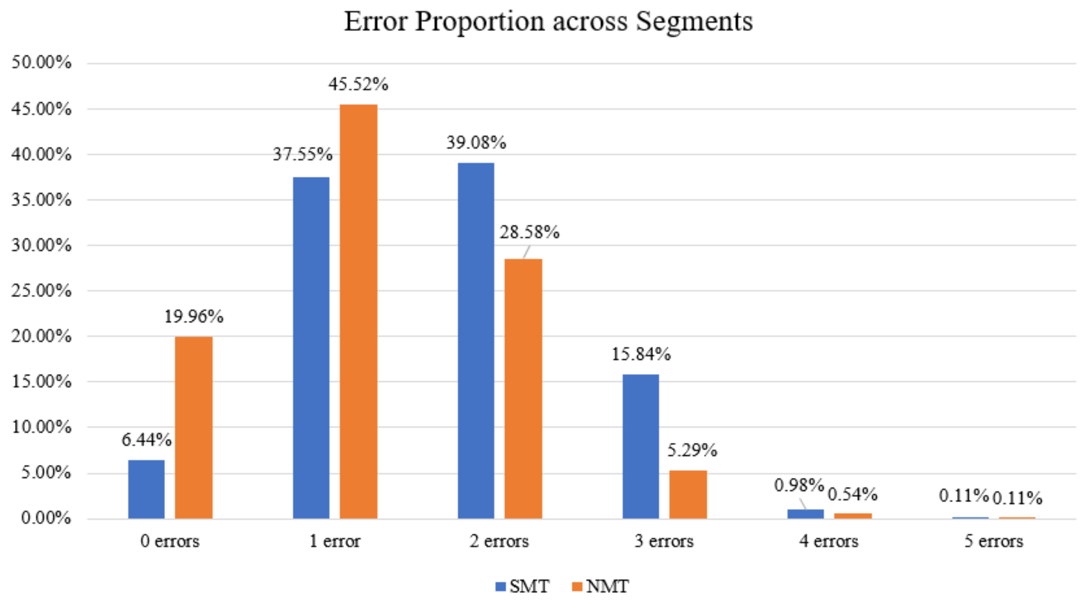

Figure 3: Error Proportion across NMT and SMT Segments in Percentages

From this point onwards, only the segments with one error in them will be used for the investigation, due to limitations in the experimental design. It should be kept in mind that even though NMT produced 422 such segments compared to only 344 by SMT, these figures do actually speak in favor of NMT, as the aim is to have as few errors within a segment as possible.

4.2 Pause Length Prior to Errors

The pause length before errors was only possible to be evaluated for segments that contain one error only. This was due to the fact that the annotation mechanism used in this study assorted the annotated errors numerically and not chronologically. This means that the pause time is not directly linked to an error annotation. Therefore, it is impossible to know which of multiple annotated errors was the first one within the segment and which pause time belongs to which editing process. Recognizing this constraint, the following analyses were conducted on a downsized dataset. SMT had a total of 344 segments containing one error and NMT a total of

422 segments. Due to this downsized data, the following results should be treated with caution as they do not represent the whole of the data.

The results of the analyses showed that contradictory to Hypothesis 2, the SMT approach required slightly less pause time prior to editing (*M* = 1.41, *SE* = .13) than the NMT approach (*M* = 1.54, *SE* = .24). Based on the distribution plots, it was obvious that the sample sizes were not normally distributed, which was also confirmed by the Kolmogorov-Smirnov test statistics for SMT *D*(344) = .303, *p* < .001 and NMT *D*(422) = .376, *p* < .001. In order to compensate for the violation of the assumption of normality, a non-parametric statistical test was conducted. A Mann-Whitney U test indicated that the pause length was not significantly greater for NMT than for SMT, *U* = 70669.50, *p* = .53.

We also conducted an analysis for the pause length prior to errors across the two translators, to see if there was a difference between them. The results of the analyses showed that Translator 1 spent more seconds pausing prior to editing (*M* = 1.85, *SE* = .31) than Translator 2 (*M* = 1.21, *SE* = .09). The assumption of normality was violated both for Translator 1 *D*(331) = .378, *p* < .001 and Translator 2 *D*(435) = .277, *p* < .001. Another Mann-Whitney U test suggested that the pause length prior to editing errors was significantly smaller for Translator 1 compared to Translator 2. Possible reasons for this are mentioned in chapter 5. As mentioned earlier, this result refers to the simplified dataset.

4.3 Pause Length Prior to Error Types

It was assumed that *Inflectional Morphology* errors would result in the smallest pauses prior to their edits, as these are generally easier errors to correct, because they only have to be corrected on the morphological level. However, results show that, in this case, *Addition* errors (i.e. something unnecessary was added to the target text) caused the shortest pause length of 1.03

seconds, followed by *Inflectional Morphology* errors with 1.35 seconds. Table 4 presents the descriptive statistics indicating the pause lengths prior to each error category.

Table 4

Descriptive Statistics of Pause Lengths Prior to Error Types.

Error Type	**Cases**	**Mean in seconds**	**Standard Error**
Inflectional Morphology	517	1.35	0.10
Word Order	49	1.41	0.21
Omission	53	1.55	0.31
Addition	20	1.03	0.13
Mistranslation	127	2.12	0.75

The assumption of normality is violated for all five error types. This was also confirmed by the Kolmogorov-Smirnov statistics for each category as can be seen in Table 5.

Table 5

Kolmogorov-Smirnov Test Results for Each Error Category.

Error Type	**KS Statistic**	**Degrees of Freedom**	**Significance**
Inflectional Morphology	.302	517	< .001
Word Order	.273	49	< .001
Omission	.303	53	< .001
Addition	.258	20	< .001
Mistranslation	.398	127	< .001

Moreover, the assumption of homogeneity of variance is violated as well. The Levene's Test of Equality of Variances yielded the following result: $F(4, 761) = 6.94$, $p < .001$. This, together with the violation of normality, leads to the fact that no parametric statistical tests can be used in order to establish significant differences between the pause times. For this reason, the non-parametric Kruskal Wallis H Test will be conducted. The assumptions for this test are met.

A Kruskal-Wallis H test showed that there was a statistically significant difference in the pause length prior to the five different error types, χ2(4) = 13.70, *p* = .008, with a mean rank pause length score of 364.93 for *Inflectional Morphology* errors, 407.70 for *Word Order* errors, 412.64 for *Omission* errors, 365.38 for *Addition* errors, 440.46 for *Mistranslation* errors. However, this does not tell us exactly what is going on in the data. It merely says that at least two of the values have a significant difference. In order to test the pairwise comparisons, we looked at the Dunn-Bonferroni post-hoc test. Tests of the 10 possible hypotheses were conducted using Bonferroni adjusted alpha levels of .005 per test (.05/10). Table 6 shows the results of the Dunn-Bonferroni post-hoc test. It shows that the significant difference yielded between *Inflectional Morphology* errors and *Mistranslation* errors.

Table 6

Pairwise Comparison of Dunn-Bonferroni Post-Hoc Test Across all Error Types.

Pairwise comparison	**Mean Rank Difference**	**Std. Error**	**Std. Test Statistic**	**Significance**
Inflectional Morph. vs. Addition	-0.45	50.43	-0.01	.99
Inflectional Morph. vs. Word Order	-42.78	33.07	-1.29	.20
Inflectional Morph. vs. Omission	-47.72	31.91	-1.50	.14
Inflectional Morph. vs. Mistranslation	-75.54	21.91	-3.45	**.001**
Addition vs. Word Order	42.33	58.71	0.72	.47
Addition vs. Omission	47.27	58.07	0.81	.42
Addition vs. Mistranslation	-75.09	53.23	-1.41	.16
Word Order vs. Omission	-4.94	43.85	-0.11	.91
Word Order vs. Mistranslation	-32.76	37.21	-0.88	.38
Omission vs. Mistranslation	-27.82	36.18	-0.77	.44

Since the data was not normally distributed, it was not possible to conduct a parametric two-way ANOVA to investigate whether there was an interaction effect between error types and translation approach. Moreover, there is no valid non-parametric test in SPSS that conducts this kind of analysis. However, there is a work-around by using the Scheirer-Ray-Hare test, which is an extension of the Kruskal-Wallis H test. The pause times of error types were ranked with ordinal values starting from 1 for the shortest pause time. With the ranked pause length data, we were able to conduct a parametric two-way ANOVA. Based on the output of the ANOVA, it was possible to calculate the Scheirer-Ray-Hare test statistics. Table 7 presents the results of this analysis. As can be seen, only the condition Error Types yielded in an H-value higher than the critical value, which represents a significant difference (*H* = 14.737, *p* = .002). There was no interaction effect between the Approach and Error Types conditions.

Table 7

ANOVA Table for Scheirer-Ray-Hare Test Results.

Source of Variation	**SS**	**Degrees of Freedom**	**MS total**	**H calculated**	**H critical p = 0.05**
Approach	4378.32	1		0.089	3.841
Error Types	719668.28	4		**14.737**	**9.488**
Interaction	164454.43	4		3.367	9.488
Total	37308030.00	764	48832.50		

In an attempt to make the most use of the data, we tried to project these results to the complete dataset. However, these calculations should be read as merely exploratory as we do not know what the numbers with a complete dataset would look like. For the purposes of this study, it is assumed that the pause time from the segments with one error can be transferred to the whole dataset. Table 8 shows the results of this estimation. This would mean that the translators would spend a total of 2,385.26 seconds (= 39.75 minutes) with pause time on SMT produced text, compared to a total of only 1,761.69 seconds (= 29.36 minutes) on NMT text. In other words, more than 10 minutes could be saved on pause time prior to error edits.

Table 8

Exploratory attempt to estimate total pause time on whole dataset.

Error Type	**Pause length per Error in Seconds**	**Errors in SMT**	**SMT - Total Pause Length in Seconds**	**Errors in NMT**	**NMT - Total Pause Length in Seconds**
Inflectional Morph.	1.35	672	907.20	562	758.70
Word Order	1.41	352	496.32	171	241.11
Omission	1.55	115	178.25	73	113.15
Addition	1.03	35	36.05	29	36.05
Mistranslation	2.12	362	767.44	289	612.68
Total		1,536	2,385.26	1,124	1,761.69

5. Discussion

The data presented in this thesis have raised some interesting points on error frequency, error types and pauses in post-editing MT content. First and foremost, it should be stated that the change of the TraMOOC platform from SMT to NMT was clearly a success. The NMT approach not only produces fewer absolute errors, it also reduces the pause length prior to editing errors during the post-editing process. In fact, NMT was able to more than triple the amount of error free segments on a sample of 1,000 segments each, which suggests a remarkable improvement. These findings are in line with previous research stating that NMT produces less errors and more fluent text (Bentivogli, Bisazza, Cettolo & Federico, 2016; Castilho et al., 2017).

The cognitive effort of certain error types is extremely hard to grasp and observe as it can be influenced by numerous individual traits of the translator that is post-editing the text (e.g. reading speed, cognitive processes, typing behavior, etc.). At first, the plan of this thesis was to investigate the cognitive load of error types based on the frameworks proposed from Komponen (2012) and Temnikova (2010). However, we realized quickly that this would not be possible with the available data as the error categories were too limited to categorize them in a sensible way into errors requiring high/low cognitive loads. Moreover, based on various data restrictions (e.g. use of only one language, use of only two translators, etc.), it would not have been possible to pinpoint the cognitive effort down to pause lengths. Instead of that, we decided to take an exploratory approach to the effect of error types on pause time in post-editing and attempted to explore if there are differences between the five annotated error types. Much like O'Brien's (2006) experiment, which attempted to find a correlation between cognitive effort and source text machine translatability, this research acknowledges that pauses in post-editing are not the only aspect of cognitive effort. It is a much more complex concept that requires further investigation.

Since pauses in post-editing and different error types have never been researched together before, we cannot compare our findings regarding the pause length to any previous research. However, we have demonstrated that there is a difference in pause length prior to post-editing across certain error types, as was initially hypothesized. The fact that the significant difference yielded between *Inflectional Morphology* errors and *Mistranslations* is in line with the assumptions from Temnikova (2010), which states that errors on the morphological level (such as *Inflectional Morphology* errors) require less cognitive effort than errors on the lexical level (such as *Mistranslations*). Nevertheless, pauses prior to editing can be influenced by various individual preferences and conditions. To mention only a few: some editors move their cursors with the mouse, others are using arrow keys. The latter naturally takes notably more time; time during which cognitive processes continue to go on, that are not recorded as pause time in this thesis. Another factor might be that some editors first cognitively go through all the errors in the sentence and then start editing, while others start editing immediately when errors are spotted.

Lacruz et al. (2012) did investigate pause time in post-editing and they categorized the segments into different cognitive loads by the number of edits that took place. They found that cognitive less demanding segments required less pause time. If we combined this with Temnikova’s (2010) theory, then our results are partially in line with these papers, as we found that *Addition* errors and *Inflectional Morphology* errors took the least pause time prior to editing the errors.

Finally, it can be said that we have re-confirmed the higher standard of NMT and we have initiated research on pauses in post-editing in relation to error types for further studies.

6. Conclusions

This thesis aims to investigate the pause length prior to different error types and to assess whether there are differences in English-German SMT and NMT generated output. Three research questions were posed prior to this thesis: *(1) Do SMT and NMT tools differ in the amount of errors that they produce in the machine translated output? (2) Do SMT and NMT tools differ in the pause length prior to editing errors? And (3) what is the effect of certain error types on the respective length of the pause that immediately precedes the errors?*

We hypothesized that NMT would both generate fewer absolute errors as well as have a smaller pause length prior to editing errors than SMT. Furthermore, an exploratory approach was taken which aimed at investigating differences in pause length prior to five error types in MT output.

This thesis is based on translations from real-life MOOC data as a part of the TraMOOC project. Four datasets with a total of 2,000 post-edited translations from English to German by two human translators were used. The results show that the NMT software generated significantly fewer absolute errors than the SMT software. Moreover, it is interesting to see that the NMT approach managed to minimize the error per segment proportion (most of the segments have only one or two errors) and to have three times as many segments without any errors than SMT. In other words, Hypothesis 1 stating that NMT would produce fewer absolute errors than SMT is confirmed.

Regarding the pause length prior to errors, the results of the analysis showed that there was no difference in pause time prior to editing errors between SMT and NMT. However, this result has to be interpreted with caution, as due to software restrictions, only the segments containing one error could be investigated. As for this thesis, Hypothesis 2 stating that NMT would require less pause time prior to error editing is rejected.

In a further step, we looked at the pause length prior to the specific error types: *Inflectional Morphology, Word Order, Omission, Addition* and *Mistranslation*. Results showed that *Inflectional Morphology* and *Addition* errors require the least amount of pause time prior to editing them. However, the sample size for *Addition* errors was extremely small compared to the other errors, which may have distorted the data. The only significant difference in error types was found between *Inflectional Morphology* errors and *Mistranslations*. Moreover, no interaction effect between the translation approaches (SMT/NMT) and the error types was found. This means that the exploratory approach in Hypothesis 3 stating that there is a difference in pause length prior to post-editing certain error types is confirmed.

One of the limitations of this thesis is the fact that only English text translated into German was investigated. The complete dataset contains translations into Portuguese, Russian and Greek as well. Therefore, in a further step it would be interesting to see whether these results hold up when applied to the other language combinations. It is assumed that the error distribution would be slightly different in a Latin or Slavic language, as *Inflectional Morphology* errors might be predominant in Germanic languages. Furthermore, only two separate translators worked on post-editing these texts. This means that many patterns that are found might be due to individual preferences from the translators. For future research attempting to investigate pause times linked to error types, it would certainly be favorable to consult more different translators to have a more diverse sample. The largest limitation and consequently also the main point for future research is concerning the annotation of error types. As stated above, we could only use segments containing one error to conduct the pause length investigations. This was due to a flawed annotation system in the design setup. Post-editors were asked to correct a segment and only when they completed the correction, they were forwarded to an annotation window where they rated the errors. This means that firstly, the annotations might not always be accurate as the translators might have forgotten exactly how many errors there were in the segments and of which error types they were, and secondly, it means that multiple error annotations were not directly linked to the correction. Hence, in the final data, two or more error types might have been annotated, but the annotations were ordered numerically and it was not possible to know which error type belonged to which correction.

Consequently, for a follow-up study we suggest an experimental design that allows the post-editors to immediately annotate the error type while they are correcting it. This way, we can be sure that all corrected errors are annotated, and we will be able to link annotation and correction together, which means we can use the complete dataset to conduct the pause length analysis.

In addition, this thesis initially planned to relate cognitive load to certain error types. However, the data did not allow to make this link as we could not be sure whether pause length was a viable indicator of cognitive load. For a follow-up study aiming to investigate the cognitive load of error types, we suggest to use a self-assessing technique by the translators with the use of Likert-type questions. Research has shown that humans are able to assess perceived cognitive load surprisingly accurately with the use of Likert-type questions (Gopher & Braune, 1984; Paas, van Merrienboer, & Adam, 1994). One idea for future research would therefore be to include a Likert-type question into every assessing window that pops up after editing an error. Another approach would be to use think-aloud protocols, which prompt the editor to comment on every decision they make and on whether they were cognitively demanding or not.

References

Allen, J. (2003). Post-editing. *Benjamins Translation Library*, *35*, 297-318.

Al-Onaizan, Y., Curin, J., Jahr, M., Knigt, K., Lafferty, J., Melamed, D. & Yarowsky, D. (1999). Statistical machine translation. In *Final Report, JHU Summer Workshop* Vol. 30.

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

Aziz, W., Castilho, S., & Specia, L. (2012). PET: A Tool for Post-editing and Assessing Machine Translation. In *LREC* 3982-3987.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109-120

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,* 263-270. Association for Computational Linguistics.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches.

Deselaers, T., Hasan, S., Bender, O., & Ney, H. (2009). A deep learning approach to machine transliteration. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 233-241. Association for Computational Linguistics.

Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., & Chen, Y. (2008). Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, 179-182. Association for Computational Linguistics.

Garcia, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation*, *25*(3), 217.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors*, *26*(5), 519-532.

Guerberof, A. (2009). Productivity and quality in MT post-editing. In *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*.

Hutchins, J., & Lovtsky, E. (2000). Petr Petrovich Troyanskii (1894–1950): A forgotten pioneer of mechanical translation. *Machine translation*, *15*(3), 187-221.

Hutchins, J. (2007). Machine translation: A concise history. *Computer aided translation: Theory and practice*, *13*, 29-70.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 181-190. Association for Computational Linguistics.

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, 11-20.

Krings, H. P., & Koby, G. S. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes* (Vol. 5). Kent State University Press.

Lacruz, I., Shreve, G. M., & Angelone, E. (2012). Average pause ratio as an indicator of cognitive effort in post-editing: A case study. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, 21-30.

Löffler-Laurian, A. M. (1986). Rapid and conventional editing: two modalities of one specific activity. *Multilingua*, *5*(4), 225-229.

McElhaney, T., & Vasconcellos, M. (1988). The translator and the postediting experience. *Technology as Translation Strategy, Binghamton, NY: State University of New York at Binghamton (SUNY)*, 140-148.

Melby, A. K., Housley, J., Fields, P. J., & Tuioti, E. (2012). Reliably assessing the quality of post-edited translation based on formalized structured translation specifications. In *IN: Proceedings of the Second Workshop on Post-Editing Technology and Practice (WPTP 2012) held at AMTA*, 31-40.

O'Brien, S. (2006). Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, *7*(1), 1-21.

Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and motor skills*, *79*(1), 419-430.

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*, 7-16.

Popovic, M., Lommel, A., Burchardt, A., Avramidis, E., & Uszkoreit, H. (2014). Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 14)*, 191-198.

Popovic, M. (2011). Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, *96*, 59-67.

Ramlow, M. (2009). *Die maschinelle Simulierbarkeit des Humanübersetzens: Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik*. Berlin: Frank & Timme GmbH.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104-3112.

Temnikova, I. P. (2010). Cognitive Evaluation Approach for a Controlled Language Post—Editing Experiment. In *LREC.*, 3485-3490.

Vilar, D., Xu, J., d'Haro, L. F., & Ney, H. (2006, May). Error analysis of statistical machine translation output. In *Proceedings of LREC*, 697-702.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Klingner, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.