universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"The Feasibility of Light Post-Editing of NMT Output in the Medical Domain for the Language Pair English-Croatian"

verfasst von / submitted by

Tina Cakara, BA BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 070 331 363
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Translation Deutsch, Bosnisch/Kroatisch/Serbisch
Betreut von / Supervisor:	Univ.-Prof Dragoş Ioan Ciobanu, PhD

Preface

I have always loved technology. I like trying out new tools and programmes, finding out how they work, understanding why they sometimes might not work and looking for a solution to fix that. However, I would have never felt confident enough with technology to choose a technological topic for my master's thesis until one online meeting that changed everything. Professor Alexandra Krause and I were having a talk about something else when out of a sudden professor Krause came up with both a rough topic idea for my master's thesis and the recommendation of professor Dragoș Ciobanu as my supervisor. On that day my master's thesis journey began.

Step by step the idea for my topic started to develop in my head. While my interest in technology is based mostly on curiosity, my love for the Croatian language goes much deeper. Growing up with Croatian I was not aware of how complex this language was. A language with seven cases that requires you to change the endings of nouns, adjectives, and pronouns only because you chose a different preposition. A language with just consonants in a word (like *čvrst*) and the declination of names (*Tina, Tine, Tini, Tinom*). A language spoken by fishermen and sailors but also by a big diaspora in the whole world. I was wondering if a language that was so challenging for me sometimes would be challenging for a machine as well. I wanted to see how well machine translation works for the language combination English-Croatian in the very specific field of medicine and how much post-editing was still needed. This is how I finally had my research question.

Like all research, this master's thesis would not have been possible without other people. First and foremost, I want to thank my supervisor Univ.-Prof. Dragoș Ioan Ciobanu, PhD for his patient, empathetic and professional supervision of my research project. His master's thesis seminar has always been a place of enthusiasm, curiosity, and intense exchange. The whole process of finding a topic, narrowing it down and formulating the research question was so much easier because of his great support. I would also like to thank Miguel Angel Rios Gaona, PhD from the HAITrans research group, who provided me with the data for my experiment and fine-tuned the NMT model mBART. With his help I could start my annotation process with a finished data set and all NMT models ready.

Finally, I would like to thank my family, friends, and partner for their motivating words during the more challenging times, for endless text messages, for chats during coffee breaks and for helpful feedback when I was not able to see the wood for the trees. Without you I would not have come so far. Thank you!

Table of Contents

1. Introduction and motivation

> If we want the best and the brightest minds on the planet working together to solve problems such as climate change, cancer, and energy crises, then we need to make sure that they can effectively share their research findings with one another. (Bowker & Ciro, 2019, p. 1)

Globalisation is shaping scholarly communication. Results are published constantly and shared with scientists all around the world. Particularly, the medical domain is characterised by intense international cooperation, leading to scientific publications mostly in English. While English is regarded as the lingua franca in everyday life as well as in scholarly communication, there are still linguistic barriers that cannot be ignored. As Bowker & Ciro (2019) put it in their book *Machine Translation and Global Research:*

> [T]hose who have only a limited command of English […] may find it challenging to produce scientific articles in English, or even search for and understand English-language publications. (p. 9)

According to them, Anglophone scientists have "the luxury and the privilege of being able to disseminate the results of their research in their own language" (p. 4). Scientists with limited or poor knowledge of English are faced with communication problems. Translation can be a solution here.

With short deadlines and a tight budget, a professional translation of scientific papers or abstracts is often not an option. In this case, machine translation (MT, and especially neural MT - NMT) can be a very useful tool to make research findings available quicker to a broader audience. On the one hand, non-Anglophone researchers can use machine translation to translate English abstracts from papers into their native languages. As Bowker & Ciro (2019) emphasise:

> [M]achine translation can be an extremely valuable tool for them [the non-Anglophone researchers] in the earlier stages of the research process, such as when they are searching for relevant research material as part of a literature survey. (p. 5)

On the other hand, Anglophone scientists can machine-translate the English abstract of their paper into other languages and let it be post-edited by non-specialist journalists before publication. This way, they can provide a summary of their research findings to non-English-speaking scientists saving time and costs, especially if the post-editing effort is not too high.

The task of post-editing (PE) is usually a necessary step when using machine translation. Independently of the different models, MT output almost always contains errors. Different guidelines have been created and published to help post-editors decide how much and what to edit. A typical categorisation of post-editing is into *light* and *full*, depending on the

number of corrections the post-editors should undertake and the expected quality of the target text.

Under these circumstances, an important research question arises, that will be investigated in this thesis: "**How applicable is the previously-published guidance for MT light post-editing (produced at the time of statistical MT - SMT) to NMT output of texts from the medical domain for the language pair English-Croatian to enable source-text understanding?"** For answering this research question, the following two sub-questions will serve as a guidance:

1) Is the level of light post-editing applied to NMT output from English into Croatian in the medical domain sufficient for content understanding, or is full post-editing required?
2) Does the existing guidance for light vs. full post-editing need to be updated to reflect NMT progress?

English abstracts from the Croatian scientific magazine *Acta Medica Croatica* will serve as source texts and their published Croatian equivalents as reference translations. 68 segments from five abstracts in English will be machine-translated by three general NMT models (Tilde, eTranslation and mBART) alongside a domain-specific fine-tuned version of mBART. The errors in the MT outputs will be annotated using the Qualitivity plug-in from Trados Studio 2021. Afterwards, the errors will be compared to guidelines for light and full post-editing by the language data network TAUS (Massaro et al., 2016), combined with criteria from other guidelines (Hu & Cadwell, 2016). Depending on the types of errors present in the NMT output for the medical domain, the existing guidance for light vs. full post-editing may need to be updated, in order to enable source text understanding.

2. Related work

2.1. The development of machine translation

The beginnings of machine translation go far back into the 1930s and 1940s. During World War II cryptographic techniques were used for codebreaking. Inspired by their success and based on these techniques, the US-American mathematician Warren Weaver, who worked for the Rockefeller Foundation, issued the so-called *Weaver's Memorandum* in 1949. This memorandum "brought the idea of machine translation to the attention of the public" (Bowker & Ciro, 2019, p. 37-38). Consequently, this led to more research, funding, discussion, and conferences about machine translation. However, after this first wave of enthusiasm, in the early 1960s it became clear

> that cryptographic techniques were too limited and that translation was far more complex than simple word-for-word substitution. (Bowker & Ciro, 2019, p. 38)

This led to the publication of a report in 1966 in the United States, issued by the Automatic Language Processing Advisory Committee (ALPAC). Bowker & Ciro (2019) summarise the process and result of the ALPAC report in the following way:

> [T]his committee examined the progress and prospects of machine translation research and concluded that it did not make sense to continue investing in machine translation. […] As a result, financial support for machine translation research virtually dried up, and the public perception of machine translation was damaged for many years to come. (p. 38)

Thus, the ALPAC report led to a drop in interest, research and funding connected to machine translation during the 1960s. However, already a few years later in the 1970s the situation changed again: "The 1970s saw some successes, particularly with systems developed to work in very specialized areas with restricted vocabularies and limited syntax." (Bowker & Ciro, 2019, p. 38)

In the 1980s the volume of text and consequently the demand for machine translations started to grow. This "created a needs-driven market and a renewed level of interest and investment in machine translation research" (p. 38). The 1990s brough with them new developments in computer-aided translations tools, for example translation memory systems (p. 38).

In the decades before the 2000s, machine translation systems were mostly rule-based (RBMT). According to Bowker & Ciro (2019), RBMT works in the following way:

> In a nutshell, researchers tried to program computers to process language in a way that resembles how human beings process language by incorporating grammar rules and large dictionaries. (p. 39)

This means that RBMT systems have to learn language rules first in order to machine translate a new piece of text. RBMT systems can be regarded as "linguistic knowledge-driven" (Brozović-Rončević et al., 2014, p. 72). They "analyse the input text and create an intermediary symbolic representation from which the target language text can be generated." (p. 72) RBMT systems strongly depend on two main factors: First, on "the availability of extensive lexicons with morphological, syntactic, and semantic information" and second, on "large sets of grammar rules carefully designed by skilled linguists" (p. 72). Both aspects require a lot of time and financial resources. At the turn of the new millennium, it came to a paradigm shift:

> [R]esearchers decided to approach machine translation in a different way, one that allowed computers to play to their strengths. (Bowker & Ciro, 2019, p. 39)

This paradigm shift led to new approaches to machine translations that were corpus-based: example-based machine translation (EBMT) and statistical machine translation (SMT). These approaches are regarded as "data-driven" (Bowker & Ciro, 2019, p. 42). In EBMT, when the system faces a new sentence, it "consults the parallel corpus to find examples of how that sentence (or at least, some of its parts) has been translated before." (p. 42) However, Bowker & Ciro (2019) point to an important aspect when using EBMT:

> [T]he parallel corpus that is provided for the machine translation system to consult must contain the right types of texts because a turn of phrase that is appropriate in one situation might not be the best choice in another. (p. 42)

This means that a parallel corpus with the same type of texts can be very useful for EBMT while a lack of the same can lead to inappropriate or useless translations.

Statistical machine translation systems are also corpus-based and data-driven. However, they work in a slightly different way than EBMT. SMT systems are also trained on parallel corpora but use probability calculations instead of examples:

> They use algorithms that give preference to sequences of words that are probable translations of source words, with a probable reordering scheme, and they generate sequences of words in the target language with high probability. (Bowker & Ciro, 2019, p. 43-44)

In other words, SMT systems rely on algorithms that calculate which translation is the most probable based on parallel corpora. According to Brozović-Rončević et al. (2014) SMT systems are "finding plausible patterns of words" (p. 72). Koehn (2020) argues that data-driven approaches like SMT "made machine translation a useful tool for many applications, from information gisting[1] to increasing the productivity of professional translators" (p. 39). However, compared to knowledge-driven MT systems like RBMT, "statistical (or data-driven) MT systems often generate ungrammatical output." (Brozović-Rončević et al., 2014, p. 72)

Both the knowledge-driven approach (RBMT) and the data-driven approach (EBMT and SMT) have advantages as well as disadvantages. A possible solution to combine the advantages of both methods is the use of hybrid approaches, as Brozović-Rončević et al. (2014) propose:

> [T]he strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. (p. 49)

In the last two decades a new approach started to emerge and brought about the neural turn: Neural machine translation "started with the integration of neural language models into traditional statistical machine translation systems" (Koehn, 2020, p. 39). Neural language models are based on neural networks, which are defined as follows:

> A neural network is a sort of information processing system that is inspired by the way that biological nervous systems, such as the brain, process information. It comprises a large number of highly interconnected processing elements that work in unison to solve specific problems. Like people, neural networks learn by example. (Bowker & Ciro, 2019, p. 44-45)

This means that the concept of neural networks derives from biology and works by creating a high number of connections between elements that get multiplied by learning from examples. According to Koehn (2020), "the entire research field of machine translation went neural" (p. 40).

[1] "[Gisting] means that a user can employ machine translation for personal use in order to get the gist or comprehend the general idea of the meaning of a text that has been written in another language" (Bowker & Ciro, 2019, p. 23).

Within only a few years the development of machine translation systems exploded:

> To give some indication of the speed of change: at the shared task for machine translation organized by WMT [=Workshop on Machine Translation], only one pure neural machine translation system was submitted in 2015. It was competitive but underperformed traditional statistical systems. A year later, in 2016, a neural machine translation system won in almost all language pairs. In 2017, almost all submissions were neural machine translation systems. (p. 40)

As this quote shows, within only three years the use of machine translation systems increased rapidly and started to become the state of art.

2.2. The use of machine translation

Using machine translation in various contexts is a common practice by now. A prominent argument speaking for the use of MT is its financial and temporal advantage. According to Way (2018),

> the range of situations in which MT is being deployed nowadays includes many where there simply is no place for human intervention, either in terms of speed, or cost, or both. (p. 2)

But before considering using machine translation, two important questions need to be asked: "how will the translation be used, and for how long will we need to consult that translation?" (Way, 2018, p. 2) The decision for or against machine translation also strongly depends on "users' needs and expectations." (Bowker & Ciro, 2019, p. 23) The expectations regarding type and quality of the translated material have been changing in the last two decades. According to Allen (2003), there is a rising need for translation gisting:

> [U]sers just want to understand in their own native language(s) the main idea(s) of a document that only exists in a foreign language (p. 300)

However, in the case of scientific abstracts, where research findings are included, the need for accuracy increases compared to an everyday use of MT. This shows that the decision for or against the use of MT needs a more complex discussion.

According to Nitzke et al. (2018) there are some key factors to consider when planning to use machine translation:

- Text type:

  > Usually, it is advisable to use MT for text types that are not very creative, contain redundancies, and may have been created using specific guidelines and rules or even using a controlled language […] Restricted texts, […] like most domain-specific communication, are, in general, more suitable for MT, because linearity and similarity to the source text is acceptable and often even favoured. (p. 243)

- Risk: “Text types that are very restricted and straightforward are often the ones that are the riskiest." (p. 243) Thus, the key factor of risk represents a contradiction to the first key factor of text type. Very high-risk texts, like warning messages, do not allow for any mistakes or misunderstandings. Therefore, they are not very suitable for MT. (p. 243)
- MT quality: “If the MT system is well-trained, the output becomes better and less effort is needed for PE.” (p. 244)
- Turnaround time and life span of translations: “Another factor that needs to be considered […] is how much time and manpower are at hand.” (p 244) If a translation will be updated soon or if the quantity of source texts is too high, a human translation might not pay off. (p. 244-245)
- Data security: “[I]t might be reasonable to assess who should have access to the MT system and it might be relevant to inform the users of the MT system about confidentiality.” (p. 245)

The decision for or against using machine translation is an ongoing discussion. Another one is concerning the question, which MT model should be used. Neural machine translation models are currently the dominant approach. But are they better compared to previous approaches, like RBMT, EBMT or SMT? Klubička et al. (2018) conducted a study of human evaluation of machine translation systems for the language pair English to Croatian. They concluded the following when comparing NMT models to phrase-based statistical MT models:

> Our NMT system produces sentences with far fewer errors, and output that is more fluent and more grammatical, which should be of help when it comes to the task of post-editing. […] NMT has shown itself to produce language that is so fluent that the fine-grained hierarchy in the Fluency branch is of little use. (p. 213)

This result strongly suggests that NMT models are performing better compared to phrase-based statistical MT systems. However, there is another side to the same story. Even though Castilho et al. (2017) found comparable results, they still emphasise:

> [E]ven though NMT shows significant improvements for some language pairs and specific domains, there is still much room for research and improvement before broad generalisations can be made (p. 110)

They go on by saying that “[e]ven though the neural model demonstrates gains in fluency, it also shows a greater number of errors of omission, addition and mistranslation.” (Castilho et al., 2017, p. 117) In their study using several language pairs (English-German, English-

Chinese, English-Greek, English-Portuguese, English-Russian), the post-editing effort was lower in NMT output compared to SMT output because of fewer morphological errors. However, according to the study participants, the NMT errors where "more difficult to identify, whereas word order errors and disfluencies requiring revision were detected faster in SMT output." (p. 117) This shows that both NMT and SMT have certain advantages depending on the needs and expectations. Way (2018) compared phrase-based statistical and neural MT models and came to a similar ambivalent conclusion:

> Our own in-house tests have shown that for small amounts of good quality training data, NMT cannot outperform PB-SMT systems. Note too that NMT currently takes much more time to train, and translation speed is slower compared to PB-SMT. (p. 13)

It is important to keep these various aspects in mind, when considering using NMT models. Coming back to the initial question, whether to use MT at all, Jia et al. (2019) conclude in their study about translation students translating and post-editing NMT output for English-Chinese:

> [T]he quality evaluation results imply that the translation products of NMT post-editing and from-scratch translation were comparable in terms of fluency and accuracy, both for domain-specific texts and general language texts. […] The results also indicate that post-editing remarkably improved the quality of the raw NMT output. (p. 79)

While this does not answer the question fully, it does show that machine translation and post-editing can reach a comparable quality in terms of fluency and accuracy as human translation. Considering this, how can machine translation be used in scholarly communication?

2.3. Machine translation literacy in scholarly communication

According to Bowker & Ciro (2019), scholarly communication can be defined as:

> the process by which academics, scholars, graduate students and other researchers share and publish their findings so that they are available to the wider research community, and beyond. As part of the scholarly communication system, knowledge is created, evaluated for quality, disseminated, and preserved for future reference. (p. 7)

In the context of scholarly communication, machine translation has emerged as a novel approach to overcome language barriers and enable communication across scientific communities. However, like with any other tool, for machine translation a certain competence is needed to correctly use it: MT literacy. Bowker & Ciro (2019) define "literacy" as "having competence or knowledge in a specified area" (p. 87).

For scholarly communication, MT literacy means that a scholar needs to:

- comprehend the basics of how machine translation systems process texts;
- understand how machine translation systems are or can be used (by oneself or by other scholars) to find, read, and/or produce scholarly publications;
- appreciate the wider implications associated with the use of machine translation;
- evaluate how (machine) translation-friendly a scholarly text is;
- create or modify a scholarly text so that it could be translated more easily by a machine translation system; and
- modify the output of a machine translation system to improve its accuracy and readability. (Bowker & Ciro, 2019, p. 88)

All these competences defining MT literacy should be part of any scholar's repertoire doing research in a globalised world. However, as Bowker & Ciro (2019) point out, "there are very few resources to help these community members acquire and teach this type of literacy." (p. 1)

Another aspect closely related to MT literacy is the use of English as a lingua franca in scholarly communication. Bowker & Ciro (2019) emphasise:

> Though English is just one of many languages spoken around the world, it is clear that it has become the leading language for scholarly publishing. (p. 9)

While English is the native language of only 6% of the world's population, the other 94% "need tools, techniques, and training to help them engage with and contribute to the scientific literature in their field" (p. 1). This shows how closely related MT literacy and English as a lingua franca are. According to Bowker & Ciro (2019), native English-speaking researchers as authors of scientific papers can support their non-native colleagues by "increasing their own level of machine translation literacy and preparing translation-friendly texts" (p. 5). They can do this by:

> learning how machine translation may affect their texts, as well as learning how to produce texts that will be more amenable to machine translation (or even simply to ensure that their untranslated texts are easier to read for those who are not native English speakers). (Bowker & Ciro 2019, p. 4)

On the other hand, non-native English speakers can use machine translation in their roles as both readers as well as authors:

> [N]on-native speakers […] may want to use machine translation to help them find and access existing scholarly literature in their field (i.e., translating from English into their native language), or may be looking to use machine translation as an aid to help them produce a text for dissemination (i.e., translating from their native language into English). (Bowker & Ciro 2019, p. 88)

Bowker & Ciro (2019) emphasise, that for researchers, free online machine translation tools are very easy to use and interact with. They just need to "select a tool, type or copy-and-paste a text, choose a language pair, and click a button" (p. 33). However, they argue that

> this does not mean that users are equipped to apply these tools successfully to their learning activities. Training in the critical and effective use of machine translation can help. (p. 33)

So, in order to improve MT literacy, Bowker & Ciro (2019) introduce a workshop divided into four modules that teaches scholars about machine translation. They conclude about MT literacy:

> [M]achine translation literacy has begun to change from a technical or specialist literacy into an everyday literacy that is starting to have implications for the way that our society communicates and goes about activities relating to research, discovery, and innovation. (p. 35)

This means that machine translation literacy is becoming an essential competence in a global and information-driven society.

2.4. The Croatian language

The Croatian language is part of the "West-South Slavic subgroup of the Slavic branch of the Indo-European linguistic family". It is the native language of around 5.5 million people worldwide (Brozović-Rončević et al., 2014, p. 50). Morphologically it is considered a language "rich in inflection" (Klubička et al., 2018, p. 200). According to the White Papers Series for Croatian (Brozović-Rončević et al., 2014), some of the morphological features that distinguish Croatian from languages like English are the following:

- five of the ten parts of speech inflect (nouns, adjectives, numbers, pronouns, verbs)
- there are three genders (masculine, feminine, neuter)
- there are seven cases (nominative, genitive, dative, accusative, vocative, locative, instrumental)
- "Verbs also have a complicated aspectual system […] and they also encode the feature of transitivity." (p. 56)

Like most other languages, Croatian has entered the digital age, being used in technological settings and by technology itself. For that to be successful, resources need to be available for a language to function in the digital world. There are languages, for which lots of resources are available. One language that is at the top of the *high-resourced* languages is English.

There are many other languages that are considered *low-resourced*. Besacier et al. (2014) define the *low-resourced languages* by the following aspects:

> lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc. (p. 87)

Other terms for the same concept are: "low-density languages, resource-poor languages, low-data languages, less-resourced languages." (p. 87) Dunđer (2020) considers Croatian a "low-resourced" language:

> As the Croatian language can be considered low-resourced in terms of available services and technology, development of new domain-specific machine translation systems is important, especially due to raised interest and needs of industry, academia and everyday users. (p. 33)

This could have two main reasons: First, the morphologically complex nature of the Croatian language above makes the development of certain language technologies more difficult than for morphologically less complex languages like English. Second, public funding in Croatia is lower than in other countries (Brozović-Rončević et al., 2014, p. 80), probably also due to the relatively low number of speakers worldwide.

However, at this point the question arises if Croatian is still a *low-resourced language*, considered that the White Paper Series was published in 2014. Even though Dunđer uses the same concept for describing Croatian in his article in 2020, the following paragraphs will give a short overview of the available resources for Croatian and their development.

The White Paper Series (Brozović-Rončević et al., 2014) gives a rather negative perspective on the availability of technological language resources for Croatian: "it can't be assumed that the current status of language technologies is satisfactory." (p. 44). An appeal is formulated:

> According to the assessment detailed in this report, focused action must be taken in order to bring the Croatian language resources and tools at the level of quality and quantity of language resources and tools that already exist for other European languages. (p. 44)

Compared to other European languages, Croatian has “weak/no support” in the area “speech processing” (see Figure 1). This includes the following:

> Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications. (Brozović-Rončević et al., 2014, p. 78-79)

English, on the contrary, has “good support”, while German has “moderate support”.

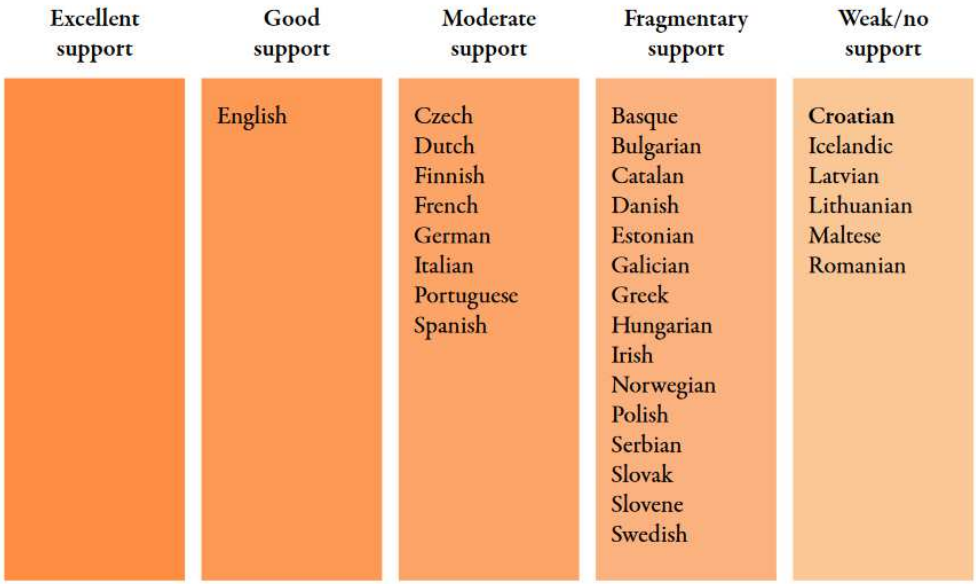

Figure 1: Speech processing: state of language technology support for 30 European languages (Brozović-Rončević et al., 2014, p. 81)

Moreover, Croatian has “weak/no support” concerning “machine translation” (see Figure 2). This includes the following aspects:

> Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications. (Brozović-Rončević et al., 2014, p. 79)

Again, English has “good support”, while German only has “fragmentary support”.

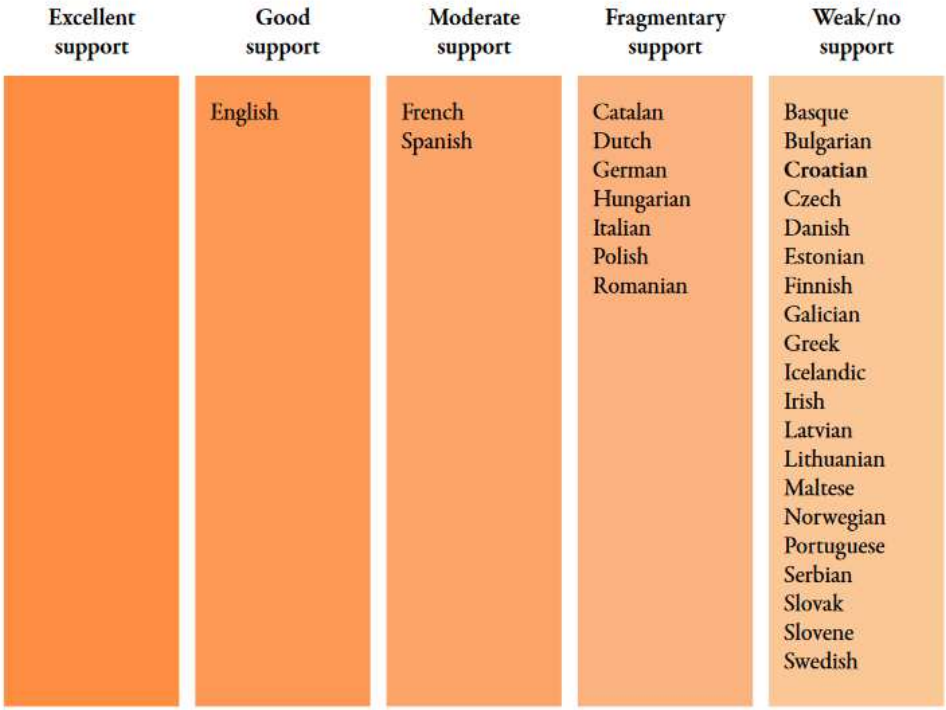

Figure 2: Machine translation: state of language technology support for 30 European languages (Brozović-Rončević et al., 2014, p. 81)

Croatian also has “weak/no support” in the field of “text analysis” (see Figure 3). This includes:

> Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars. (Brozović-Rončević et al., 2014, p. 79)

Once more, English has “good support”, while German has “moderate support”.

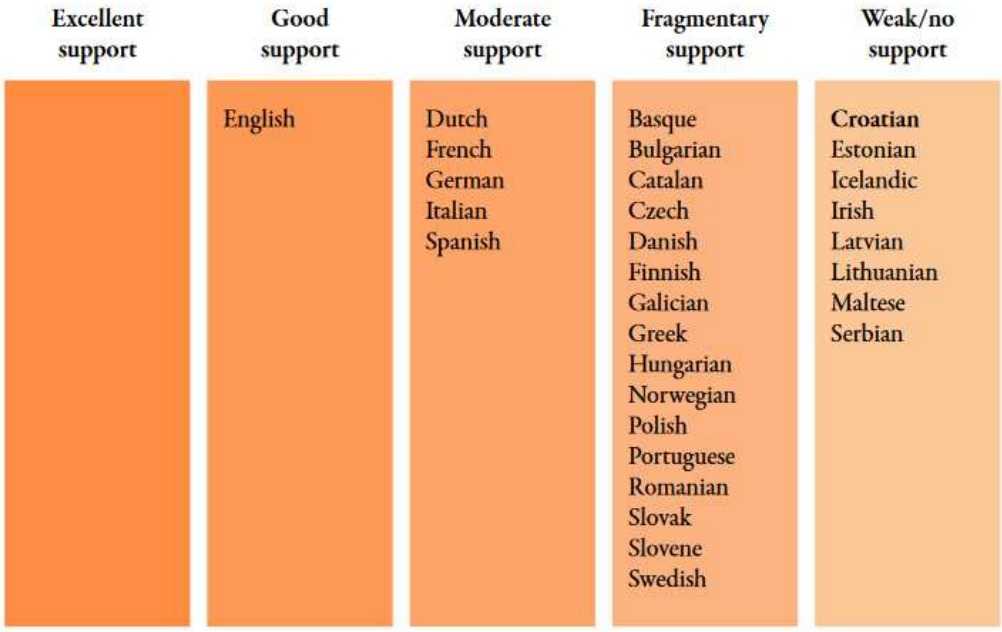

Figure 3: Text analysis: state of language technology support for 30 European languages (Brozović-Rončević et al., 2014, p. 82)

Finally, Croatian has "fragmentary support" in the area "speech and text resources" (see Figure 4). This includes the following:

> Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars. (Brozović-Rončević et al., 2014, p. 79)

Again, English has "good support" as the only language, while German has "modest support".

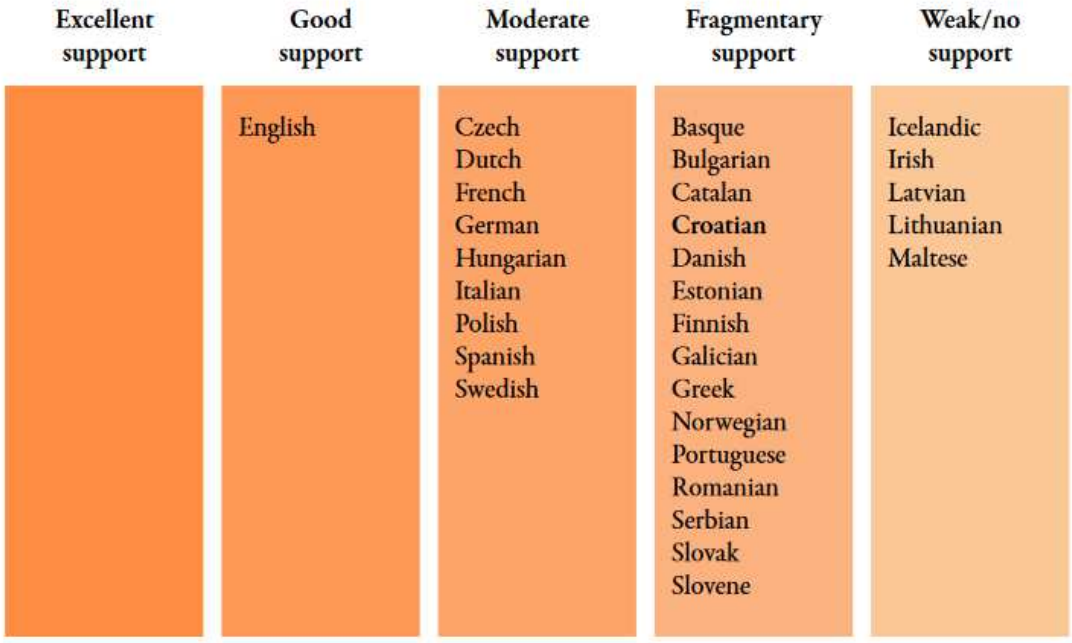

Figure 4: Speech and text resources: state of language technology support for 30 European languages (Brozović-Rončević et al., 2014, p. 82)

However, in the White Paper Series (Brozović-Rončević et al., 2014), some projects are mentioned that are aiming to improve the availability of technological language resources for the Croatian language. In the lexical area, machine-readable thesauri and ontological language resources are being developed:

> [They] have demonstrated improvements in finding pages using synonyms of the original search terms, such as nuklearna energija and atomska energija (nuclear energy and atomic energy) or even more loosely related terms. (p. 68)

Another project is the Croatian Lemmatisation Server[2], developed at the University of Zagreb. Its goal is "to be able to search for all the inflectional forms of a word at once, instead of having to enter each different form separately." (p. 69) Concerning machine translation systems, the project "Information Technology in Translation and e-Learning of Croatian" was initiated in 2007 with the goal to "investigate the prerequisites in building MT systems

[2] See http://hml.ffzg.hr/hml/info.php?show=hlp, retrieved 6 November 2022.

for translation into and from Croatian." (Brozović-Rončević et al., 2014, p. 73) To sum up, the Croatian language is represented by basic language technology:

> Croatian stands reasonably well with respect to the most basic language technology tools and resources, such as reference corpora, smaller parallel corpora, large inflectional lexicons, tokenisers, MSD taggers, lemmatisers, NERC system etc. (Brozović-Rončević et al., 2014, p. 77)

Some of these basic resources for the Croatian language, which have been researched and found for this project, are the following:

- Croatian Language Corpus[3]
- Croatian Special Field Terminology[4]
- Croatian Terminology Portal[5]
- Croatian National Corpus[6]
- Croatian Morphological Lexicon[7]
- Croatian Morphological Database of Verbs[8]
- Croatian Dependency Treebank[9]

On the website of Sketch Engine[10] a list of tools to work with Croatian text corpora is given (Croatian Word Sketch, Croatian concordance, Croatian term extraction, Bilingual term extraction, Croatian thesaurus, Croatian word lists, N-grams in Croatian), as well as a list of Croatian corpora.

Despite these resources for the Croatian language, a few missing resources for Croatian are mentioned in the White Paper Series:

> Tools and resources for more advanced language technology such as deep parsing, machine translation, text semantics, discourse processing, language generation, dialogue management, etc., simply do not exist. (Brozović-Rončević et al., 2014, p. 77)

Again, it needs to be mentioned that the White Paper Series has been published in 2014, just a year after Croatia has become part of the European Union. This step, as well as huge improvements in technology during the last decade have changed the availability of technological language resources for the Croatian language a lot. Some of the mentioned areas in the previous citation could therefore be partly challenged:

---

[3] http://riznica.ihjj.hr/, retrieved 6 November 2022.
[4] http://struna.ihjj.hr/en/, retrieved 6 November 2022.
[5] http://nazivlje.hr/english/, retrieved 6 November 2022.
[6] http://filip.ffzg.hr/cgi-bin/run.cgi/first_form, retrieved 6 November 2022.
[7] http://hml.ffzg.hr/hml//?lang=en, retrieved 6 November 2022.
[8] http://croderiv.ffzg.hr/Croderiv//, retrieved 6 November 2022.
[9] http://hobs.ffzg.hr/en/search///, retrieved 6 November 2022.
[10] https://www.sketchengine.eu/corpora-and-languages/croatian-text-corpora/, retrieved 6 November 2022.

Concerning deep parsing there exists the project CLASSLA pipeline[11] as well as Universal Dependencies formalism[12] for a similar process, namely dependency parsing. Machine translation has improved, as it is stated in the White Paper Series as well:

> Google Translate has offered translations to and from Croatian since 2008. The quality of the translations was rather poor in the beginning, but is getting better as more and more parallel Croatian-English data is available on-line. (Brozović-Rončević et al., 2014, p. 73)

This aspect is essential for this project, as the results comparing the different NMT models in this thesis will show. Concerning discourse processing, the dataset FRENK[13] includes data about offensive language discourse. Even for language generation, a new project for Croatian automatic text generation has been started[14], trained on Croatian Wikipedia content.

All these examples show that more and more resources are being developed for Croatian, making it less of a *low-resourced language* than during the time when the White Paper Series was published. Some of the most recent technological language resources for Croatian can be found on the Slovenian page of CLARIN[15] (Common Language Resources and Technology Infrastructure). They offer an FAQ (Frequently asked questions) section, covering the following aspects: online Croatian language resources, tools to annotate Croatian texts, and datasets to train Croatian annotation tools.

2.5. Post-editing

According to the ISO 17100 (2015), post-editing means to "edit and correct machine translation output" (cited by Hu & Cadwell, 2016, p. 346). Bowker & Ciro (2019) go into more detail when defining post-editing:

> [P]ost-editing typically refers to a situation where a language professional takes the raw machine translation output and corrects any errors in order to bring it up to an acceptable quality. Indeed, post-editing can be carried out to various levels depending on user needs. (p. 25)

This quote emphasises that post-editing should be done by language professionals and its goal is acceptable quality. This acceptable quality depends on the user's need and can be achieved with various levels of post-editing. Allen (2003) points out:

> It is therefore important to determine to what extent MT output texts are acceptable, and how much human effort is necessary to improve such imperfect texts. (p. 298)

[11] https://github.com/clarinsi/classla, retrieved 6 November 2022.
[12] https://universaldependencies.org/u/overview/syntax.html, retrieved 6 November 2022.
[13] https://www.clarin.si/repository/xmlui/handle/11356/1462, retrieved 6 November 2022.
[14] https://huggingface.co/macedonizer/hr-gpt2, retrieved 6 November 2022.
[15] https://www.clarin.si/info/k-centre/faq4croatian/, retrieved 6 November 2022.

What both quotes do not suggest is that the post-editor should also look into the source content during the post-editing process, thus giving a misleading picture of what post-editing is. The monolingual perspective on post-editing will be discussed in more detail in section 5.3. In general, the raw MT output quality and the post-editing effort vary depending on the following factors: MT system, language pair and direction, domain, text type, and degree of control over the input text (O'Brien, 2010, p. 8).

A common classification of post-editing effort is into *light* and *full* post-editing (Hu & Cadwell, 2016; Nitzke et al., 2018; O'Brien, 2010). While *light* post-editing ensures that the "quality is good enough or understandable", in *full* post-editing a "human-like" quality is the aim (Hu & Cadwell, 2016, p. 347). O'Brien (2010) mentions other commonly used terms for the two levels of post-editing:

- Light: Fast PE, Gist PE, Rapid PE
- Full: Conventional PE (p. 3-4)

Typically, light post-editing is "for internal dissemination", while full post-editing is "for wide dissemination or certified documentation" (Hu & Cadwell, 2016, p. 347). Allen (2003) uses a similar classification, differentiating between *inbound* and *outbound* translation:

> Inbound translation (also referred to elsewhere as MT for acquisition or assimilation) is simply the process of translating to understand. Outbound translation (also referred to as MT for dissemination), on the other hand, the process of translating to communicate. (p. 301)

In scholarly communication, both inbound and outbound translation are relevant. On the one hand, from the perspective of a non-Anglophone scientist, trying to understand a paper or abstract published in English would fall under inbound translation. On the other hand, from the perspective of an Anglophone scientist, translating their English papers or abstracts to communicate their research to other scientists would count as outbound translation.

The question arises, how the exact criteria for light and full post-editing can be determined. In the industry, guidelines for post-editing are used. However, very few of them have been released online (Hu & Cadwell, 2016, p. 348). Nunziatini and Marg (2020) point to the following problems with post-editing guidelines:

> While existing definitions of light and full post-editing are useful as general guidelines, they typically remain too abstract and inflexible both for translation buyers and linguists. Besides, they are inconsistent and overlap across the literature and different Language Service Providers (LSPs). (p. 1)

On the one hand, the guidelines are expressed in a very vague way and can be unclear to translators, clients, and scientists themselves. On the other hand, there are no standard guidelines that are consistent across different Language Service Providers. This can lead to confusion for post-editors which can result in "editing too much (over-editing) or not enough (under-editing)" (Nunziatini & Marg, 2020, p. 6).

Nitzke et al. (2018) introduce a helpful decision tree model (see Figure 5) for estimating if light or post-editing is required or if the raw MT output is sufficient. They explain who the target audience for that model is:

> [T]his decision model is described from the customer's point of view and/or the person who decides if MT will be used and what degree of PE effort is necessary. (p. 246)

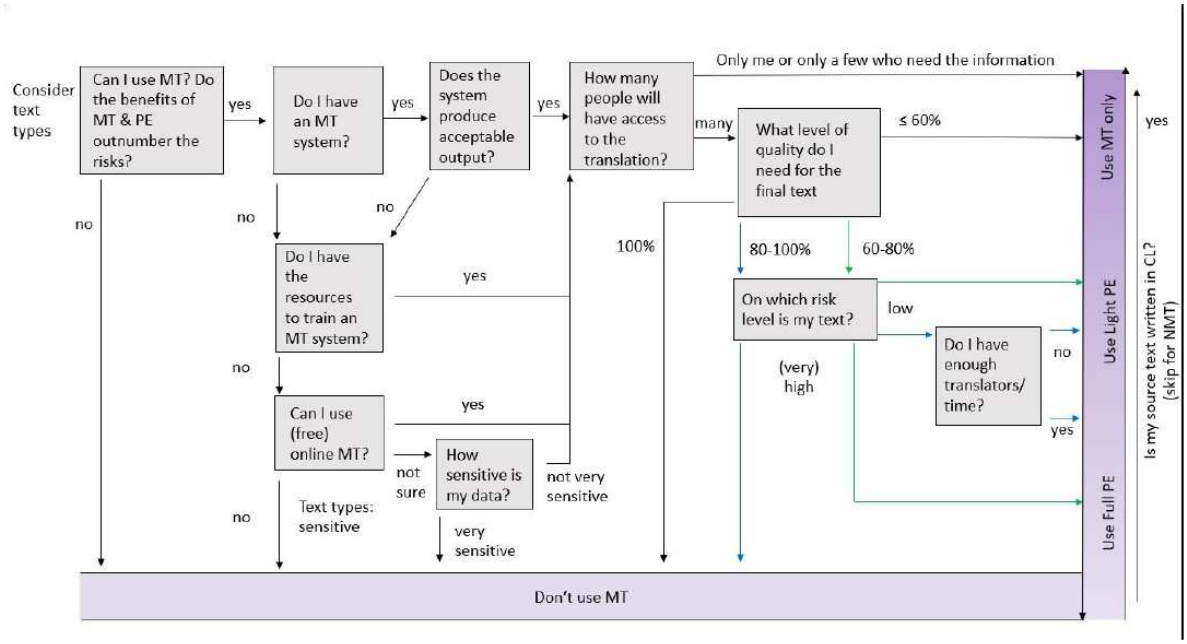

Figure 5: Decision tree model for PE tasks (Nitzke et al., 2018, p. 246)

However, this decision tree model leaves important questions regarding quality open: What does 60%, 80% or 100% quality actually mean? What is "acceptable output" and is it the same for all users? And does human translation always lead to 100% accuracy and quality? The assumption that human translation provides maximum quality and can always be used as a gold standard should be viewed critically. There is evidence that human translations do not always offer 100% quality, even after revision (Ciobanu et al., 2019; Daems & Macken, 2020). This will be discussed in more detail in section 5.3., based on the error analysis of the reference translations used for this project.

According to Nunziatini & Marg (2020), even just the terms *light* and *full* post-editing can lead to confusion, especially for clients, because

> often people misunderstand that these describe how much editing needs to be done, or in other words, how much effort the post-editor should put into the task, rather than what the final translation quality should be. More precisely, some people might erroneously think that, if they were to translate the same content in multiple languages, depending on the quality of the raw MT output, some languages will require light post-editing while others will require full post-editing. (p. 3-4)

So, on the one hand, clients might not be aware that light and full post-editing describe the quality of the final translation rather than the process of editing. On the other hand, they might believe that a language like English, for example, needs "only" light post-editing, while a morphologically more complex language like Croatian requires full post-editing. Nevertheless, the source language is very important when using machine translation, even despite the fact that nowadays more and more NMT models are multilingual. The myths about machine translation need to be addressed in order to minimise confusion for clients. This is very important because clients "are themselves often not entirely sure which approach would meet, exceed or fall short of their requirements." (Nunziatini & Marg, 2020, p.1)

In order to meet even more needs, Chung-ling Shih (2021) presents a triangular set of PE guidelines, setting a counterpoint to the dual approach. The three goals of post-editing are: "amending linguistic, pragmatic and affective MT errors" (p. 125). The author lists specific strategies to achieve those three goals. For amending linguistic MT errors, post-editors should apply the following strategies:

> (a) use of words with correct meanings based on the context, (b) adaptation of metaphorical expressions, (c) adaptation of cultural references, (d) use of correct grammar, and (e) use of correct word order. (p. 140)

In order to fix pragmatic MT errors, the following strategies are recommended:

> (f) elimination of redundant, mechanical and repetitive words, (g) use of consistent words, (h) supplementation of additional words or information, (i) rewriting or paraphrasing of the entire clause or sentence, and (j) use of well established, register-specific terms of the target language. (p. 144)

Finally, to edit affective MT errors, the author suggests the following strategies:

> (k) rewriting the sentence with a new thematic focus, and (l) using catchy description. (p. 148)

This triangular approach by Chung-ling Shih offers a new and innovative perspective on post-editing, based upon examples for the language pair English-Chinese. However, these guidelines are not a common practise in the language industry yet. They should function as

a side note to show the variety of approaches to post-editing and making it more precise and client-oriented.

2.6. Error annotation

During the process of post-editing, errors in MT output are identified and corrected. To evaluate MT output and/or compare different MT engines, errors can be annotated using an error typology. According to Esperança-Rodier (2020), who conducted a survey on the language pair English-French, "the choice of a typology impacts the results." (p. 2) So, depending on the chosen error typology, error occurrences as well as error types might vary (p. 7). There are two main challenges when annotating errors:

> A recurring discrepancy is the selection issue: difference in the number of annotations depending on the typology used. The second recurring issue is the annotation issue: for the same error, the annotator annotated it under one of the typologies and not the other one. (Esperança-Rodier, 2020, p. 4)

The selection and the annotation issue need to be kept in mind when deciding on an error typology for annotating errors. Another issue not mentioned in the quote by Esperança-Rodier is the problem of annotation procedures: some approaches stop after annotating one error, while others annotate all the errors in that segment. Most typologies, like the MQM-DQF quality framework (Nunziatini & Marg, 2020, p. 2), classify the errors according to two main categories: *accuracy* and *fluency*. Van Brussel et al. (2018) define *accuracy* as:

> the transfer of information and meaning from source to target language. The following main categories can be distinguished: mistranslation, do-not-translate (DNT), untranslated, addition, omission, and mechanical. (p. 3801)

Compared to that, *fluency*:

> deals with the well-formedness of the target language, regardless of the transmission of content and meaning from the source into the target sentence. The following categories are identified: grammar, lexicon, orthography, multiple errors and other. (Van Brussel et al., 2018, p. 3802)

There is a crucial difference between these two major error categories. Klubička et al. (2018) summarise it in the following way:

> [F]luency issues can be assessed without regard to whether the text is a translation or not. So for example, if a translated text tells the user to push a button when the source tells the user not to push it, there is an accuracy issue, while a spelling error or a problem with register remain issues regardless of whether the text is translated. (p. 203)

However, this seemingly clear-cut distinction between *accuracy* and *fluency* errors is not always applicable in real-life situations. In this case, it is up to the annotator to decide for the right category, as Klubička et al. (2018) argue:

> Very often examples can seem like they belong to either category, and so it is up to the annotators' judgement to decide which level is a better fit, and then being consistent in following through on the decisions made regarding dubious examples. (p. 203)

This research project is based on the MQM error typology (Multidimensional Quality Metrics), which will be described in the following paragraphs. The typology was created in the context of the projects QTLaunchPad and QT21, which were funded by the European Union, and it was directed by the German Research Center for Artificial Intelligence (DFKI). Initially it was based on the LISA QA Model but changed to a more modular approach now (MQM Council, 2022a). The MQM error typology went through different stages of development:

> […] TAUS and DFKI, working with LTAC Global, created a DQF:MQM, a harmonized error typology based on the first versions of MQM and the TAUS Dynamic Quality Framework (DQF). The DQF error typology became a subset of MQM. This has since been superseded by MQM-Core, which updates DQF:MQM and replaces it. (MQM Council, 2022a)

This means that MQM was merged together with DQF into the error typology DQF:MQM where DQF was a subset of MQM. Then MQM-Core started to update and slowly replace DQF:MQM. Today, MQM is maintained and monitored by the MQM Council (MQM Council, 2022a).

In short, MQM is a "methodology for analytic Translation Quality Evaluation (TQE)" (MQM Council, 2022b). Its goal is to provide "a standardized way to generate structured information about the quality of translated content" (MQM Council, 2022b). MQM assists translation evaluators in the following aspects:

- Measuring the quality of a translation product
- Identifying specific errors in translated content that need to be corrected
- Improving communication concerning expectations, quality requirements and performance
- Enabling data-driven quality management (MQM Council, 2022b)

Regarding the first aspect, measuring the quality of a translation product, MQM helps identify issues regarding quality in the whole text, as well as specific errors. This shows which quality requirements have been met and if a translation can be accepted, needs to be revised, or must be rejected (MQM Council, 2022b).

For the second aspect, identifying specific errors in translated content that need to be corrected, MQM enables the evaluator to "determine the impact of the detected issues on the translation product's suitability for its intended purpose". This impact is categorised into different "severity levels" (MQM Council, 2022b).

Concerning the third aspect, improving communication concerning expectations, quality requirements and performance, MQM reduces subjectivity and ambiguity and therefore "greatly increases the efficiency of cooperation between the stakeholders involved" (MQM Council, 2022b).

Finally, for the fourth aspect, enabling data-driven quality management, MQM provides structured evaluation data that helps improve quality in a systematic and continuous way. The source content and whole translation process can be improved through MQM's "error analysis and root cause analysis" (MQM Council, 2022b).

Compared to BLEU, an automatic evaluation metric for machine translation, MQM is "a manual, reference-free approach that identifies specific translation errors". In contrast, BLEU needs reference translations to build its scores (MQM Council, 2022c). For this research project MQM is used even though reference translations are available. The reason behind this is because the goal of this project is not to evaluate different NMT models according to their BLEU score but to identify specific errors in the NMT outputs that need to be corrected (see second aspect of MQM mentioned above). This will then help to answer the research question regarding the feasibility of light post-editing for NMT output.

3. Methodology

The goal of this master's thesis is to examine the following research question: "How applicable is the previously-published guidance for MT light post-editing (produced at the time of statistical MT - SMT) to NMT output of texts from the medical domain for the language pair English-Croatian to enable source-text understanding?" For answering this research question, the following two sub-questions will serve as a guidance:

1) Is the level of light post-editing applied to NMT output from English into Croatian in the medical domain sufficient for content understanding, or is full post-editing required?
2) Does the existing guidance for light vs. full post-editing need to be updated to reflect NMT progress?

To investigate the research question, the output of three general NMT models (Tilde, eTranslation and mBART) was analysed alongside a domain-specific fine-tuned version of mBART according to accuracy and fluency against existing reference translations (RTs). English abstracts from the Croatian scientific magazine *Acta Medica Croatica* served as source texts and their published Croatian equivalents as reference translations.

In a first step, all errors in the different versions of NMT output were annotated using the Qualitivity plug-in of Trados Studio 2021 compared to the RTs. In case of doubt about some errors, validation with a subject matter expert from Croatia helped clarify the issues.

In a second step, the error categories found were compared to guidelines for light and full post-editing by TAUS (Massaro et al., 2016) combined with criteria from other guidelines (Hu & Cadwell, 2016).

The two steps of this research project follow O'Brien's (2010) ways of measuring quality for post-editing. The first step corresponds to her point "Types of errors: Compares source text with raw MT output" (p. 10). By comparing the raw MT output against the source text, as well as the reference translations, the diverse types of errors were identified. The second step is related to her point "Changes made: Compares post-edited text with raw MT output" (p. 10). By comparing the type and severity of annotated errors for the raw MT output against the reference translation, the errors were classified according to the guidelines for light vs. full post-editing. This approach combines a quantitative and a qualitative method. The empirical data collected with error annotation in step 1 is quantitative, but the comparison of the error categories found in the MT output with the post-editing guidelines

and the analysis in step 2 is of qualitative nature. The results will thus indicate whether current guidance for light post-editing needs to be updated for the medical domain to match the development of NMT.

3.1. Dataset

Scientific abstracts from the Croatian medical journal *Acta Medica Croatica* (Vol. 75, No. 2 and 3, 2021) authored in English were used as source texts (68 segments from five abstracts in total), and their published Croatian equivalents as reference translations. Initially it was a challenge to find English source texts that have Croatian reference translations because that is an uncommon language direction. Most articles in the medical journals of Acta Medica were either authored in Croatian with English abstracts or in English with no Croatian abstract. Only a few articles were authored in both languages. The goal for this project was to find English abstracts with Croatian reference translations. Therefore, the articles with English abstracts needed to be authored in English as well. Five English abstracts with Croatian reference translations (68 segments in total) were chosen for the final data set.

English as a source language was chosen because of its dominance as a lingua franca in scholarly communication (see section 2.3), while Croatian as a target language was chosen because of its linguistic contrast to English. As Klubička et al. (2018) point out for their survey on the language pair English-Croatian: "English-to-Croatian [is] a language direction that involves translating into a morphologically rich language" (p. 195) They list some linguistic examples of this exact language pair:

> The fact that Croatian is rich in inflection, has rather free word order and other similar phenomena not present in English gives rise to specific translation issues. For example, grammatical categories that do not exist in English, like gender or case inflections in nouns, may be particularly hard to generate reliably in a Croatian translation. (Klubička et al., 2018, p. 200)

This morphological difference between English and Croatian makes it more challenging for NMT models to provide correct translations.

The medical domain was chosen due to three reasons: First, like it was mentioned before, the medical domain is shaped by intense international cooperation, resulting in lots of communication across language barriers. Machine translation as a tool can be very useful in this situation. Second, the medical domain is a highly specialised field dominated by technical language and terminology. This has an impact on machine translatability. And third, for this research project, the multilingual NMT model mBART was fine-tuned for the medical domain by Miguel Angel Rios Gaona, PhD from the HAITrans research group

(https://haitrans.univie.ac.at/team/). To fine-tune NMT models, bilingual domain-specific corpora are necessary. However, as Nitzke (2018) points out: "Still, for some language pairs none or only very little data might be available. Hence, it might be problematic or even impossible to train a system." (p. 244) Luckily, for the medical domain a multilingual corpus in English and Croatian made out of PDF documents from the European Medicines Agency (EMEA) was available (Prokopidis & Papavassiliou, 2020) and it could therefore be used for fine-tuning mBART.

The English source texts were selected according to two criteria: First, it was checked if there was a Croatian reference translation of the English abstract. This was important in order to use the reference translations as a *gold standard* for annotating the NMT outputs. If the reference translations could really be used as a gold standard will be discussed in more detail in section 5.3. Second, it was checked if the English abstracts and their Croatian reference translations were equivalents in a way that they could be used as aligned parallel texts. If they differed to much because whole parts were left out or added, they were not chosen for this project.

3.2. NMT models

The NMT models used for this research project are three general ones, Tilde (*EU Council Presidency Translator*, n.d.), eTranslation (*Presidency MT*., n.d.) and mBART (Tang et al., 2020), as well as a fine-tuned version of mBART (Tang et al., 2020) for the medical domain.

The first two general NMT models, Tilde (*EU Council Presidency Translator*, n.d.) and eTranslation (*Presidency MT*., n.d.), are both implementations of the EU Council Presidency Translator, which was created for the Croatian Presidency of the Council of the European Union in the year 2020. On its website, the creation and the aim of the machine translation tool are explained:

> The EU Council Presidency Translator is a multilingual communication tool that enables delegates, journalists, translators, and visitors to cross language barriers and access information [...] To provide more fluent translations for the local languages of the hosting countries in 2020, the tool also features AI-powered Neural MT systems. When translating, Neural MT systems examine the full context of a sentence, producing highly readable translations that are almost human-like in style. The EU Council Presidency Translator was designed and developed by Tilde, a language technology company that specializes in Neural MT, in close partnership with Croatia and the Faculty of Humanities and Social Sciences of the University of Zagreb support from the CEF eTranslation building block. (*EU Council Presidency Translator*, n.d.)

The difference between Tilde and eTranslation is that Tilde develops its own general, custom or adaptive MT systems according to different needs, whereas eTranslation is a NMT model

developed by and for the needs of translators at the EU (Tilde Senior Account Manager, personal communication, October 27, 2022).

The third general NMT model mBART (Tang et al., 2020) works in a slightly different way than the first two NMT models:

> MBART is a multilingual encoder-decoder (sequence-to-sequence) model primarily intended for translation task. As the model is multilingual it expects the sequences in a different format. A special language id token is added in both the source and target text. (von Platen, n.d.)

Finally, a fine-tuned version of mBART (Tang et al., 2020) is used as a fourth NMT model for this research. It was fine-tuned by Miguel Angel Rios Gaona, PhD from the HAITrans research group for the medical domain with a multilingual corpus (English-Croatian) made out of PDF documents from the European Medicines Agency (EMEA) (Prokopidis & Papavassiliou, 2020).

3.3. Error annotation with Qualitivity

In this research project, the errors in the output of the NMT engines, as well as errors in the reference translations were annotated with the help of Qualitivity, a plug-in for Trados Studio 2021. It offers a variety of features:

> In short, it provides functionality to track the time spent on translating, reviewing & post-editing the segments from documents, but additionally includes functionality to track every single change made to the segments at a granular level and a means to generate reports based on that data in structured and readable format. (*Qualitivity*, n.d.)

The name "Qualitivity" itself "suggests that it is a cross between productivity and quality, simply because that is the type of functionality that it is equipped to provide." (*Qualitivity*, n.d.) One feature is particularly relevant for this project, namely the *Quality Metrics*: "this allows the user to import some of the more known standards such as TAUS DQF, SAE J2450 & MQM Core... for measuring quality." (*Qualitivity*, n.d.) A Quality Assessment report can then be generated and exported into XML or an Excel file. Depending on the source text type the user is evaluating, the reports can differ as well: "[Qualitivity] might also be applicable in assessing the translations from an MT engine in a certain domain" (*Qualitivity*, n.d.). The last feature was used in this research project.

For the quality metrics the standard from MQM - Multidimensional Quality Metrics (MQM Council, 2022d) was imported into the Qualitivity plug-in of Trados Studio 2021 together with a separate set of categories for the terminology errors by Haque et al. (2019).

The following categories were imported for error annotation (their definitions can be found in the appendix):

Accuracy

Accuracy > Mistranslation

Accuracy > Omission

Accuracy > Addition

Accuracy > Terminology

Accuracy > Terminology > Reorder error

Accuracy > Terminology > Inflectional error

Accuracy > Terminology > Partial error

Accuracy > Terminology > Incorrect lexical selection

Accuracy > Terminology > Term drop

Accuracy > Terminology > Source term copied

Accuracy > Terminology > Disambiguation issue in target

Accuracy > Terminology > Other error

Accuracy > Untranslated

Fluency

Fluency > Content > Inconsistency

Fluency > Content > Register

Fluency > Content > Stylistics

Fluency > Mechanical > Grammar

Fluency > Mechanical > Locale convention

Fluency > Mechanical > Spelling

Fluency > Mechanical > Style guide

Fluency > Mechanical > Typography

Fluency > Unintelligible

Verity > Completeness

Verity > Legal requirements

Verity > Locale-specific content

Some error categories above are marked in orange because there was no error in these categories in neither the reference translations nor any of the NMT outputs. Therefore, they are left out later in the figures in chapters 4 and 5 showing the findings. However, they are included here for completeness.

MQM offers four levels of error severity. For this paper the three levels *Critical, Major* and *Minor* were used based on the definitions by Lommel (2018):

- **Critical**:

  > Critical errors are those that by themselves render a project unfit for purpose. Even a single critical error would prevent a translation from fulfilling its purpose (e.g. by preventing the intended user from completing a task) and may have safety or legal implications. (p. 120)

- **Major**:

  > Major errors make the intended meaning of the text unclear in such a way that the intended user cannot recover the meaning from the text, but are unlikely to cause harm. They must be fixed before release, but if they were not they would not result in negative outcomes (other than possibly annoyance for the user). (p. 120)

- **Minor**:

  > Minor errors are those that do not impact usability. In most cases the intended user will correct them and move on, perhaps without even noticing them. (p. 120-121)

- **Null**:

  > The null level is used to mark changes that are not errors. For example, if the requestor decides to change a term after a translation is submitted, the reviewer could mark Terminology issues with this severity. No penalties are applied at this severity. It was added to MQM in 2014 to improve compatibility with the TAUS DQF. (p. 121)

The level *Null* was left out because it only marks "changes that are not errors", which would only apply to the reference translations in this project. However, this would make the comparison between the reference translations and the NMT outputs more difficult. Therefore, all changes were annotated at least as a *Minor* error. This does not have an influence on the results because the values of the severity levels (10 for *Critical,* 5 for *Major,* 1 for *Minor,* 0 for *Null*) were not added together. Only their absolute number is important for this project.

In the following paragraphs the annotation process with Qualitivity will be described in more detail. First, the Qualitivity plug-in was installed in Trados Studio 2021 and the error categories mentioned above (MQM and terminology error categories by Haque et al. 2019) were imported. Then the reference translations and the NMT outputs in Croatian were aligned with the source texts in English and imported as bilingual Excel files into Trados Studio 2021. The annotation process was carried out in parallel for the reference translation

and all outputs. This means that the errors in the first segment were annotated for the reference translations and all NMT outputs before moving on to the next segment. This ensured consistency between the annotated errors and created comparable data. Every error was assigned an error category, as well as a severity level. After finishing the annotation process, a Qualitivity report was exported.

Figure 6 shows an extract of the Qualitivity report with the most important information marked in colour.

53	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	203	8ca11ee9-1	bd4f8092-07aec251-1	Accuracy > Terminology > Incorrect le	Open	Major	5	klinike protiv rakova.	should be "antirabične
53	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	203	8ca11ee9-1	580c300e-cac7a050-1	Fluency > Mechanical > Grammar	Open	Minor	1	Zagrebske	wrong declination, sho
55	EN source_HR eTranslation.xlsx.sdlxliff	a14f07ea-1	EN source	210	c43131b5-1	06749a46-60277884-1	Accuracy > Terminology > Other error	Open	Minor	1	antibjesne klinike	should be "antirabične
56	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	214	406e326e-1	3851a4ef-2512dafe-1	Accuracy > Terminology	Open	Minor	1	Rezultati	Source said "data"
56	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	214	406e326e-1	515afd00-06793ac7-1	Accuracy > Omission	Open	Minor	1	službenog registra	"patient" was not trans
57	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	218	4fd745b8-1	74cfa4d1-d417c2af-1	Fluency > Mechanical > Typography	Open	Minor	1	%	spaces
57	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	218	4fd745b8-1	e4d74121-b648b5ec-1	Fluency > Mechanical > Grammar	Open	Minor	1	su bile	wrong word order
57	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	218	4fd745b8-1	fac89493-10501ad2-1	Accuracy > Mistranslation	Open	Major	5	os	
58	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	220	8ca11ee9-1	2436e575-5cf3febc-1	Fluency > Mechanical > Grammar	Open	Minor	1	najčešće vrste, zabilježene	should be singular
60	EN source_HR eTranslation.xlsx.sdlxliff	a14f07ea-1	EN source	227	c43131b5-1	3693cb1e-0005e58e-1	Fluency > Mechanical > Grammar	Open	Minor	1	najučestalije vrste, zabilježene	should be singular
60	EN source_HR eTranslation.xlsx.sdlxliff	a14f07ea-1	EN source	227	c43131b5-1	eb142244-66ef8ad2-1	Fluency > Mechanical > Typography	Open	Minor	1	%	spaces
60	EN source_HR eTranslation.xlsx.sdlxliff	a14f07ea-1	EN source	227	c43131b5-1	ec08f47a-f26a5cfe-1	Fluency > Mechanical > Grammar	Open	Minor	1	svinje	should be singular
61	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	228	406e326e-1	b42d5ee4-4dc5f7da-1	Fluency > Mechanical > Typography	Open	Minor	1	%	spaces
65	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	230	8ca11ee9-1	20b079e7-dbd4d415-1	Accuracy > Mistranslation	Open	Major	5	skloniji ozljedama	
65	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	230	8ca11ee9-1	cc83e150-e9f576e1-1	Fluency > Mechanical > Grammar	Open	Minor	1	u udovima ili rukama i prstima.	wrong declination
67	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	233	406e326e-1	2f931561-20728d67-1	Accuracy > Terminology	Open	Major	5	dijagnosticiranih	it does not say that the
67	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	233	406e326e-1	56e55169-e3a65e14-1	Accuracy > Omission	Open	Minor	1	znakove	"microscopic" was not
70	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	260	8ca11ee9-1	0af42a66-1cb7a948-1	Fluency > Mechanical > Grammar	Open	Minor	1	upalu	wrong declination
71	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	272	4fd745b8-1	927dd7ae-f4670b93-1	Fluency > Mechanical > Grammar	Open	Minor	1	kliničkі	wrong declination
71	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	272	4fd745b8-1	083a869e-0ff582b1-1	Fluency > Mechanical > Spelling	Open	Minor	1	appendicitis	
72	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	281	4fd745b8-1	f515a9ed-de7a268b-1	Accuracy > Terminology > Partial error	Open	Major	5	vaskularni prijelomi	
9	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	3	406e326e-1	836eb4f1-e2bda73d-1	Accuracy > Omission	Open	Minor	1	,	"and" was not translate
19	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	34	8ca11ee9-1	8079eb11-fb7deea5-1	Fluency > Mechanical > Grammar	Open	Minor	1	je sadržavala	word order
78	EN source_HR reference.xlsx.sdlxliff	51447a36-1	EN source	342	406e326e-1	92ccaff9-caaa0059c-1	Fluency > Mechanical > Spelling	Open	Minor	1	appendicitis	
79	EN source_HR eTranslation.xlsx.sdlxliff	a14f07ea-1	EN source	343	c43131b5-1	8670d357-d7ad7169-1	Accuracy > Mistranslation	Open	Major	5	pogoršati	it says the opposite
81	EN source_HR mBART general.xlsx.sdlx	1336ee52-1	EN source	347	8ca11ee9-1	5b6831dd-ac58d206-1	Fluency > Mechanical > Grammar	Open	Minor	1	praksu	wrong declination
82	EN source_HR mBART fine tuned.xlsx.s	617b352a-1	EN source	349	4fd745b8-1	a798ecdb-b00f8c09-1	Fluency > Mechanical > Spelling	Open	Minor	1	apendiktomije	

Figure 6: Report exported from the Qualitivity plug-in in Trados Studio 2021

The name of the output in the report is marked in green. In this example, “HR eTranslation” is the output of the NMT model eTranslation. The error category is marked in red, in this case “Fluency > Mechanical > Grammar” with the different category levels divided by the sign “>”. The error severity is marked in blue, “Minor” in this example. The content (word or phrase) that contained the error is marked in yellow and the comment of the annotator is marked in purple. Any changes made to the annotated data afterwards were done directly in the report. This mostly included the change of the error severity after comparing a particular error with similar errors in other outputs. The finalised report was then used for the analysis.

For analysing the data, different filters were applied in the report. The filtered data was used to create different charts to visualise and present the findings for every output separately, for comparisons of certain outputs (Tilde vs. eTranslation, mBART general vs. mBART fine-tuned) and for comparisons of all outputs (see chapter 4). The annotator’s comments (marked in purple above) were used as a basis for the examples of errors in certain error categories.

3.4. Post-editing guidelines

After annotating the errors in all the NMT outputs, they were compared to guidelines for light vs. full post-editing. However, as mentioned before, "there are no widely accepted general or standard post-editing (PE) guidelines" (Hu & Cadwell, 2016, p. 346). This means that in order to have a basis for comparing the errors in the NMT outputs, first the various guidelines that exist need to be discussed.

One set of common post-editing guidelines was published by the language data network TAUS (Massaro et al., 2016). In 2010, their guidelines were the "first attempt at publicly available industry-focused PE guidelines" (Hu & Cadwell, 2016, p. 347). In 2016, they updated the guidelines to include more elements (p. 347). TAUS emphasises that these guidelines are not fixed for all organisations and contexts, but vary according to different parameters: "We expect that organisations will use these basic guidelines and will tailor them as required for their own purposes." (Massaro et al., 2016, p. 14) According to TAUS, an ideal post-editing situation would look like this:

> Generally, these guidelines assume bilingual post-editing (not monolingual) ideally carried out by a paid translator, but that might in some scenarios be carried out by bilingual domain experts or volunteers. The guidelines are not system or language-specific. (Massaro et al., 2016, p. 14)

The effort involved in post-editing is determined by two main criteria: First, the quality of the MT raw output and second, the expected end quality of the content (Massaro et al., 2016, p. 16). While TAUS also uses the terms *light* and *full* post-editing, they decided to differentiate the two levels of end quality rather than the post-editing process itself. The two levels of end quality that TAUS uses are: "'good enough' or 'fit for purpose'" quality and "'high-quality human translation and revision' (a.k.a. 'publishable quality')" (Massaro et al., 2016, p. 16). "Good enough" quality is defined in the following way:

> [C]omprehensible (i.e. you can understand the main content of the message), accurate (i.e. it communicates the same meaning as the source text), without being stylistically compelling. The text may sound like it was generated by a computer, syntax might be somewhat unusual, grammar may not be perfect, but the message is accurate. (Massaro et al., 2016, p. 17)

TAUS presents specific guidelines for achieving "good enough" quality:

- Aim for semantically correct translation.
- 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 apply.

- No need to implement corrections that are of a stylistic nature only.
- No need to restructure sentences solely to improve the natural flow of the text. (Massaro et al., 2016, p. 17)

Compared to that, "human translation quality" is defined as:

> [C]omprehensible (i.e. an end user perfectly understands the content of the message), accurate (i.e. it communicates the same meaning as the source text), stylistically fine, though the style may not be as good as that achieved by a native-speaking human translator. Syntax is normal, grammar and punctuation are correct. (Massaro et al., 2016, p. 18)

Specific guidelines for achieving "human translation quality" are the following:

- 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. (Massaro et al., 2016, p. 18)

In their study, Hu and Cadwell (2016) compared the guidelines by TAUS (Massaro et al., 2016) with guidelines by O'Brien (2010), Mesa-Lao (2013) and Densmer (2014). They put all proposals for light vs. full post-editing into two tables (see Figure 7 and Figure 8 on the next pages), that I will use as a basis for comparing the NMT output in my study.

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

Figure 7: Comparative study of light PE guidelines (Hu & Cadwell, 2016, p. 349)

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

Figure 8: Comparative study of full PE guidelines (Hu & Cadwell, 2016, p. 350)

Hu and Cadwell (2016) summarised the results of their comparison of the four different guidelines:

> [T]he existing PE guidelines have many overlaps, especially for light post-editing. The main differences lie in the full PE guidelines and concern the requirement for style and the expected quality of the target text[.] (p. 351)

For this research project, the TAUS guidelines were combined with the comparison by Hu and Cadwell (2016, pp. 349–350) to create a new collection of guidelines for light and full post-editing. The aim was to include as many of the mentioned aspects as possible to create a clearer picture of what a post-editor should correct when requested to do light or full post-editing. The guidelines that were used for this research project are summarised in Table 1.

Table 1: Guidelines for light and full post-editing

CRITERIA	LIGHT POST-EDITING	FULL POST-EDITING
Accuracy	TT communicates the same meaning as ST (TAUS 2016). Factually accurate (Densmer 2014).	TT communicates the same meaning as ST (TAUS 2016). Absolutely accurate (Densmer 2014)
Semantics	Aim for semantically correct translation (TAUS 2016).	Aim for […] semantically correct translation (TAUS 2016).
Grammar	May not be perfect (TAUS 2016).	Aim for grammatically […] correct translation (TAUS 2016).
Syntax	Might be unusual (TAUS 2016).	Aim for […] syntactically […] correct translation (TAUS 2016).
Terminology	No need to spend too much time researching if incorrect (Mesa-Lao 2013). Be consistent (Densmer 2014).	Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms (TAUS 2016).
Information	Ensure that no information has been accidentally added or omitted (TAUS 2016).	Ensure that no information has been accidentally added or omitted (TAUS 2016).

Culture	Edit any offensive, inappropriate or culturally unacceptable content (TAUS 2016).	Edit any offensive, inappropriate or culturally unacceptable content. Adapt all cultural references (Densmer 2014).
MT output	Use as much of the raw MT output as possible (TAUS 2016).	Use as much of the raw MT output as possible (TAUS 2016).
Spelling	Basic rules regarding spelling apply (TAUS 2016).	Basic rules regarding spelling […] apply (TAUS 2016).
Punctuation	-	Basic rules regarding […] punctuation and hyphenation apply (TAUS 2016).
Format	-	Ensure that formatting is correct. All tags are present and in the correct positions (O'Brien 2010).
Style	No need to implement corrections that are of a stylistic nature only (TAUS 2016).	Consistent, appropriate and fluent (Densmer 2014).
Restructure	No need to restructure sentences solely to improve the natural flow of the text (TAUS 2016). Rewrite confusing sentences (Densmer 2014).	Rewrite confusing sentences (Densmer 2014).

The guidelines in Table 1 were used to identify the types of annotated errors in the output of the three NMT models as well as the reference translations. The aim was to answer the following questions regarding NMT output in the medical domain for the language pair English-Croatian:

1) Is the level of light post-editing applied to NMT output from English into Croatian in the medical domain sufficient for content understanding, or is full post-editing required?
2) Does the existing guidance for light vs. full post-editing need to be updated to reflect NMT progress?

4. Findings and analysis

4.1. Accuracy errors

Accuracy errors concern "the transfer of information and meaning from source to target language." (Van Brussel et al., 2018, p. 3801) For this project a total of 12 error categories was considered when annotating *Accuracy* errors in the reference translations, as well as the outputs of the four NMT models. Figure 9 depicts the findings of *Accuracy* errors in the reference translations. The graph shows the absolute number of errors in each error category as well as the proportion of errors in a particular error category to the total number of errors. In all the reference translations (68 segments from five medical abstracts), a total number of 45 *Accuracy* errors was found.

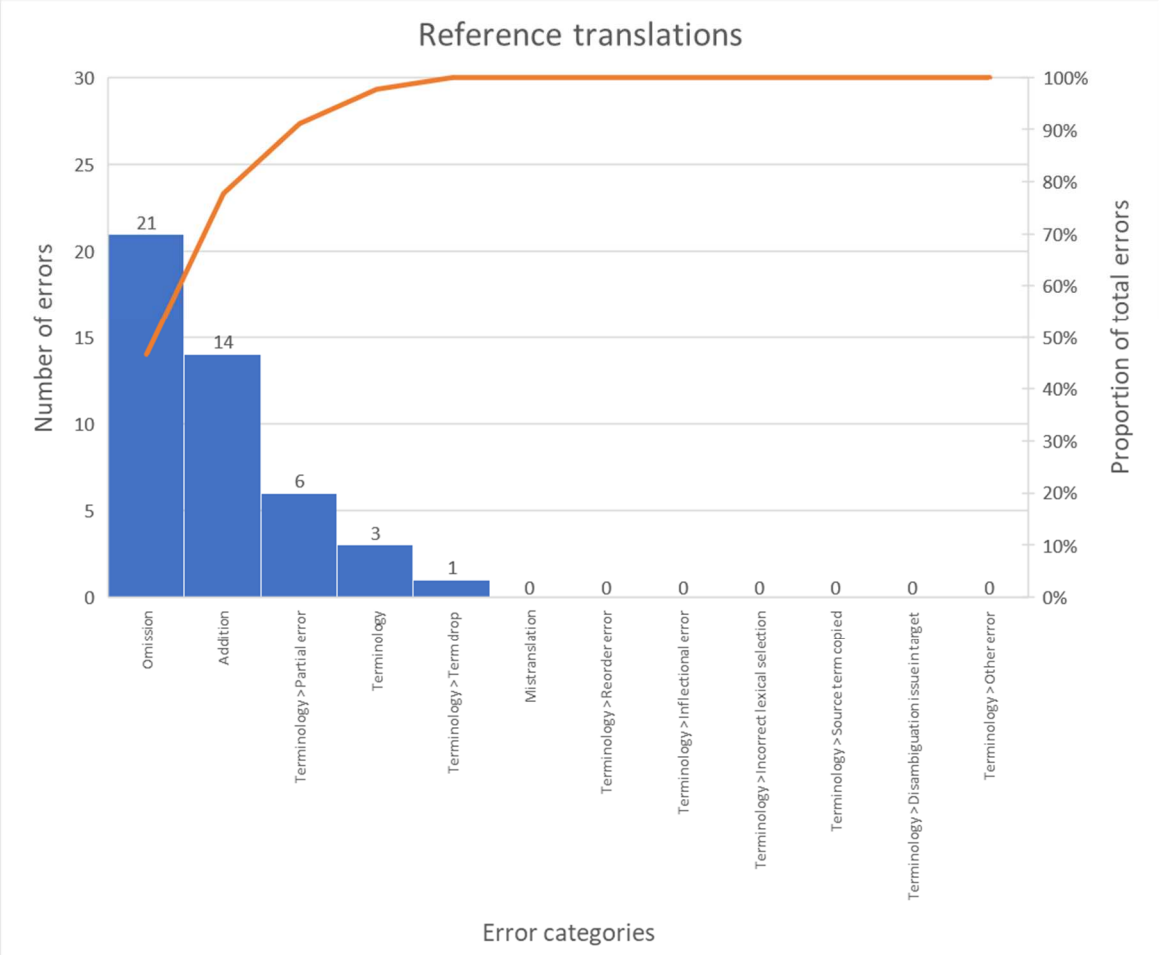

Figure 9: Accuracy errors in reference translations

The frequency of the errors in the error categories decreases from left to right. The most frequent error category in the reference translations is *Omission* with 21 errors, which makes up around 47% of all errors in the reference translations (see orange line). The *Omission*

errors varied from *Minor* errors to *Major* errors. An example of a *Minor Omission* error is the following: "after his initial hospitalization"[16] in English was translated as "nakon hospitalizacije", into Croatian leaving out the word "initial". While this example still conveys most of the meaning of the source text, the following example of a *Major Omission* error leaves out crucial information:

> [it] showed a small right-sided pneumothorax and pneumomediastinum **along the bronchi, large blood vessels, and cardiac contour** with 'ground-glass' opacifications in all lung lobes. (English source text, my emphasis)
>
> bio je vidljiv manji desnostrani pneumotoraks te pneumomedijastinum uz uzorak takozvanog 'mliječnog stakla' u svim plućnim režnjevima. (Croatian reference translation)

The highlighted part in the English source text was omitted in the Croatian reference translation. This could be either by accident or because it was assumed that some of the information is known to the readers beforehand or is made clear by the rest of the abstract.

The second frequent error category found in the reference translations is *Addition* with 14 errors, which together with the first category *Omission* makes up around 78% of all errors. An example of a *Minor Addition* error is the translation of the subtitle "Methods" in English into "Uzorak i metode" in Croatian. Thus, the word "uzorak" (meaning "specimen" in English) was added, probably for more precision. There were no *Major Addition* errors in the reference translations.

The third frequent error category is *Terminology > Partial error* with six errors. An example of this is the translation of "control group" in English into "kontrola" in Croatian (*Major* error). Here, the term was only partially translated into Croatian. The translation of the word "group" ("skupina" in Croatian) is missing, which would have resulted in the correct Croatian term "kontrolna skupina". Together, the three most frequent error categories make up around 91% of all errors. While there are only three errors in the category *Terminology* and one error in the category *Terminology > Term drop*, there are no errors in the six remaining categories.

[16] All examples cited in the following chapter are taken from the dataset used for this project and will not be quoted from here. The source texts of the dataset were taken from the medical journal *Acta Medica Croatica* (Vol. 75, No. 2 and 3, 2021). The target texts were produced by the NMT models Tilde, eTranslation, mBART and mBART fine-tuned.

Figure 10 shows the findings of *Accuracy* errors in the output of the NMT model Tilde. A total number of 50 *Accuracy* errors was found.

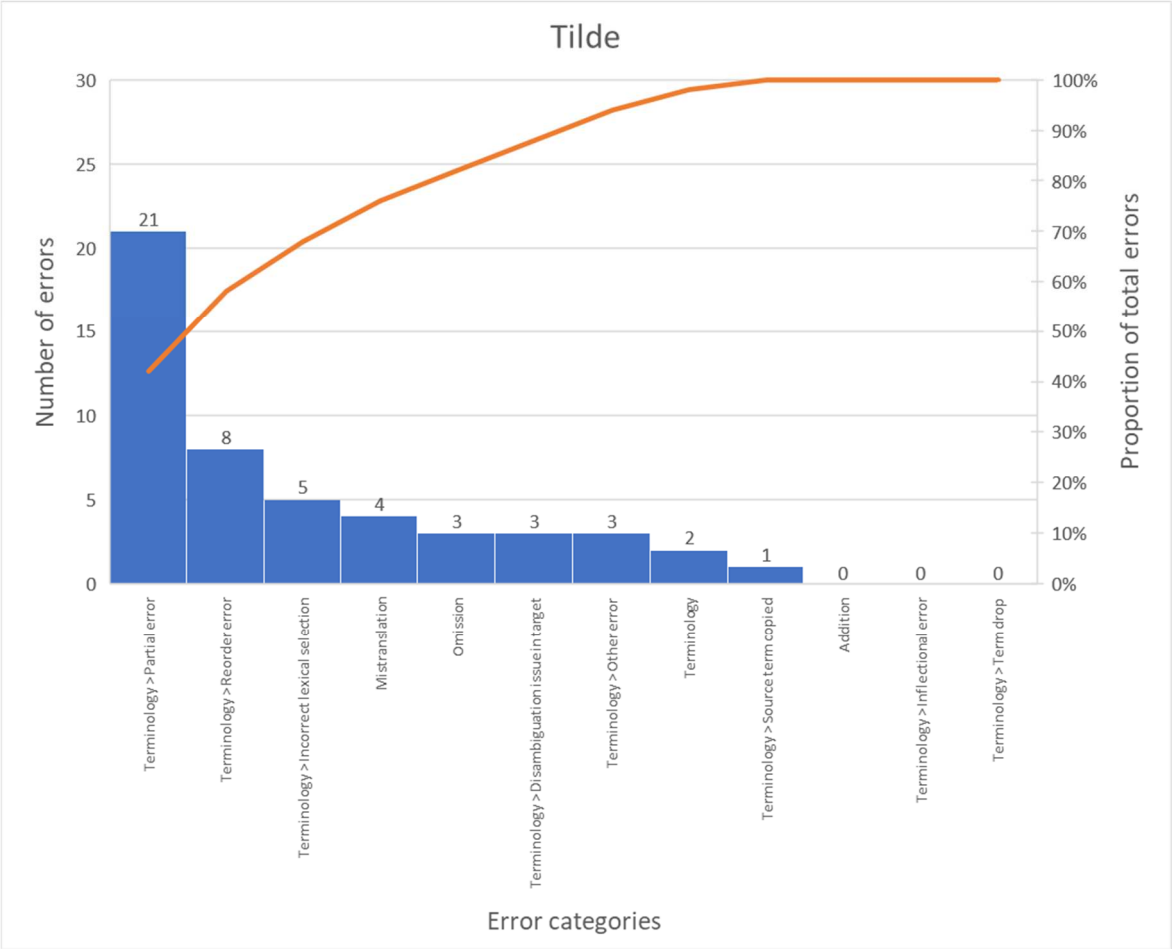

Figure 10: Accuracy errors in Tilde output

The most frequent error category in the output of Tilde is *Terminology > Partial error.* While it has more than the double number of errors (21 errors) compared to the second frequent category (8 errors), it still makes up only 42% of all errors in the Tilde output, similar to the reference translations (around 47% for the most frequent category). An example of a *Minor Terminology > Partial error* is the following: "members of the Croatian Association for the Treatment of Pain" in English was translated into "članovi Hrvatskog **saveza** za liječenje boli" in Croatian (my emphasis), whereas it should be "članovi Hrvatskoga **društva** za liječenje boli" (my emphasis). Another example of a *Major Terminology > Partial error* is the translation of the term "COVID-19" in English into "ko21-19" in Croatian. While the second part "-19" is translated correctly, the first part "ko21" is a combination of a mistranslation ("ko") and an addition ("21").

The second frequent error category is *Terminology > Reorder error* with eight errors. Together with the most frequent category it makes up 58% of all errors in the Tilde output. An example of a *Major* error in this error category is the following: "chronic pain treatment" in English was translated into "kroničnо liječenje boli" in Croatian, which would mean something like "chronic treatment of pain" in English. Thus, the reordering of the words changed the meaning of the content. The correct translation into Croatian would be "liječenje kronične boli".

The third frequent error category in the Tilde output is *Terminology > Incorrect lexical selection* with five errors. An example of a *Major* error is the translation of "CATP members" in English (Croatian Association for the Treatment of Pain) into "Članovi CAPA" in Croatian. In this case, the abbreviation "CAPA" is incorrect. The correct translation would be "Članovi HDLB" in Croatian (Hrvatsko društvo za liječenje boli). This is a *Major* error because the abbreviation (which is not the same as the English one) on its own does not mean anything to a Croatian reader. While in the reference translation, the three most frequent error categories make up around 91% of all errors, in the Tilde output the three most frequent error categories make up only 68%. There are no errors in three categories.

Figure 11 shows the findings of *Accuracy* errors in the output of the NMT model eTranslation, where a total number of 64 *Accuracy* errors was found.

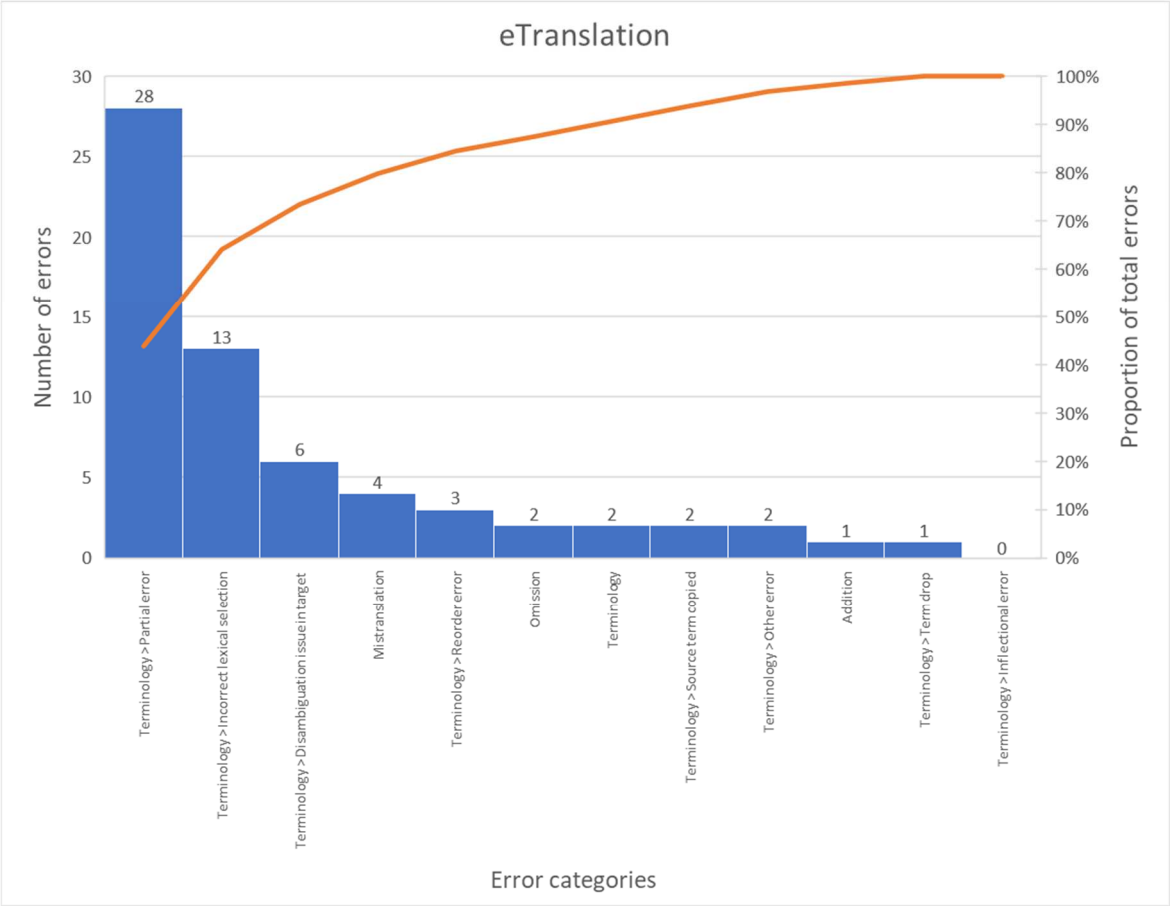

Figure 11: Accuracy errors in eTranslation output

The most frequent error category in the output of eTranslation is *Terminology > Partial error* with 28 errors, making up around 44% of the total amount of errors in the output. An example of a *Major Terminology > Partial error* is the following: the term "vascular clefts" in English was translated into "vaskularni čepovi" in Croatian, whereas the correct translation would be "vaskularni kleftovi". This means that "vascular" in English was translated correctly while the translation of "clefts" was incorrect.

The second frequent error category is *Terminology > Incorrect lexical selection* with 13 errors. Together with the first category, it makes up around 64% of all errors. An example of a *Major* error in this category is the translation of "acute appendicitis" in English into "akutni dodatak" in Croatian (the meaning of "dodatak" is closer to the English "addition" or "supplement"), whereas the correct translation would be "akutni appendicitis".

The third frequent error category is *Terminology > Disambiguation issue in target* with six errors. An example of a *Major* error in this category is the translation of the English

term “fecaliths” into “fekalije” in Croatian (meaning “faeces” in English), which completely alters the meaning of the source term. The correct Croatian translation would be “fekoliti”. There is a similar number of errors in the other categories. While the most frequent three categories make up around 73% of all errors, there are no errors in the last category.

Figure 12 shows the findings of *Accuracy* errors in the output of the NMT model mBART general. A total number of 75 *Accuracy* errors was found.

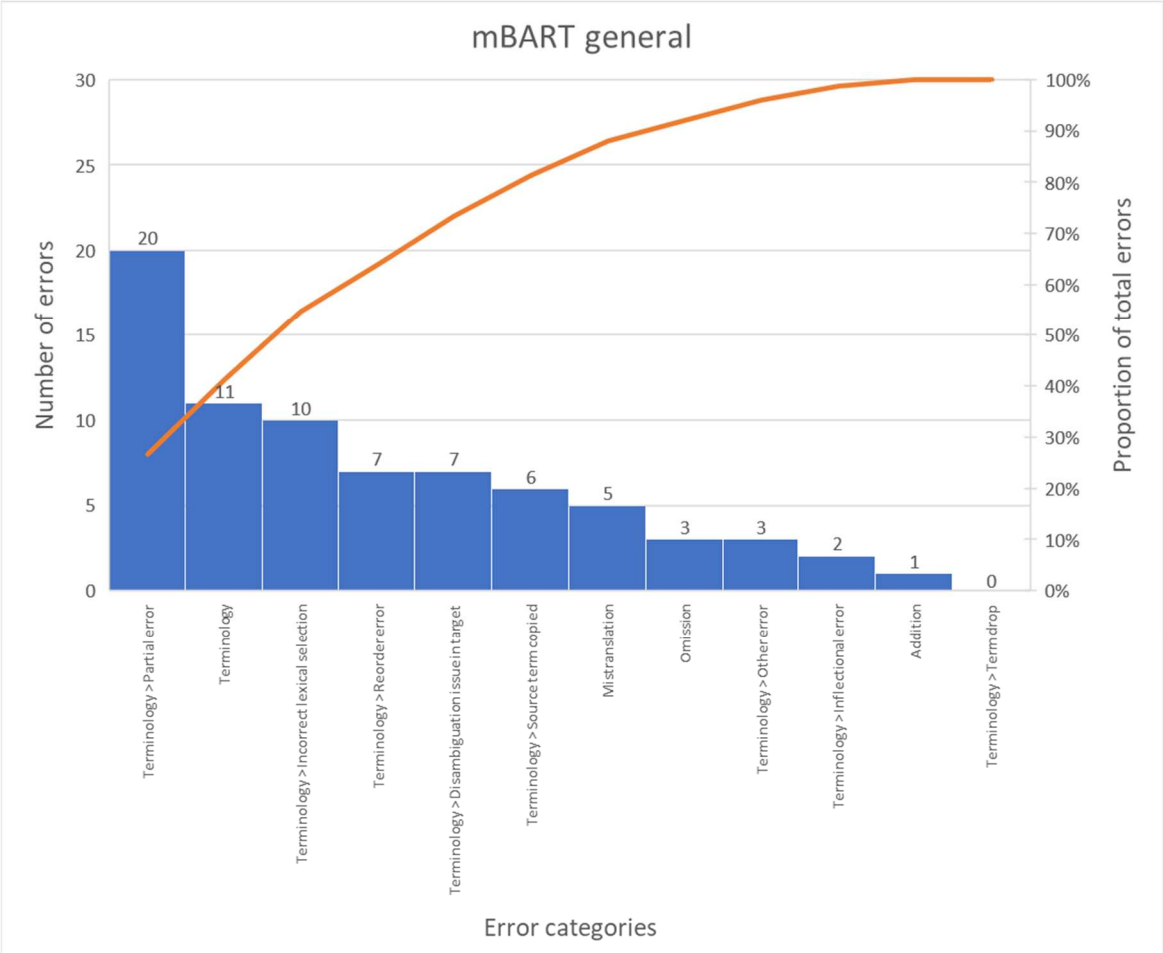

Figure 12: Accuracy errors in mBART general output

The most frequent error category in the output of mBART general is *Terminology > Partial error* with 20 errors. Even though it has almost more than the double number of errors (20 errors) compared to the second frequent category (11 errors), it makes up only 27% of the total amount of errors. This is much less compared to the reference translations (47%), Tilde (42%) and eTranslation (44%). An example of a *Major* error in the category *Terminology > Partial error* is the translation of “Croatian Association for the Treatment of Pain” in English into “chorvatske organizacije za liječenje boli” in Croatian, whereas the correct translation

would be “Hrvatsko društvo za liječenje boli”. The first part of the name of the organisation was translated incorrectly, while the second part is correct.

The second frequent error category is *Terminology* in general with 11 errors. Together with the first error category, it makes up around 48% of the total amount of errors in the mBART general output. An example of a *Minor Terminology* error is the translation of “long-term consequences” in English into “duge posljedice” in Croatian (meaning something like “long consequences” in English), whereas “dugoročne posljedice” would be the correct translation.

The third frequent error category is *Terminology > Incorrect lexical selection* with ten errors (only one less than in the previous category). An example of a *Major* error in this category is the following: “serum irisin measurement” in English was translated into “mjerenje irisina u cjelini” in Croatian (“cjelina” meaning something like “entirety” in English). The correct translation would be “mjerenje irisina u serumu”. The most frequent three error categories make up around 55% of all errors. There is no error in the last category.

Figure 13 shows the findings of *Accuracy* errors in the output of the NMT model mBART fine-tuned, where a total number of 54 *Accuracy* errors was found.

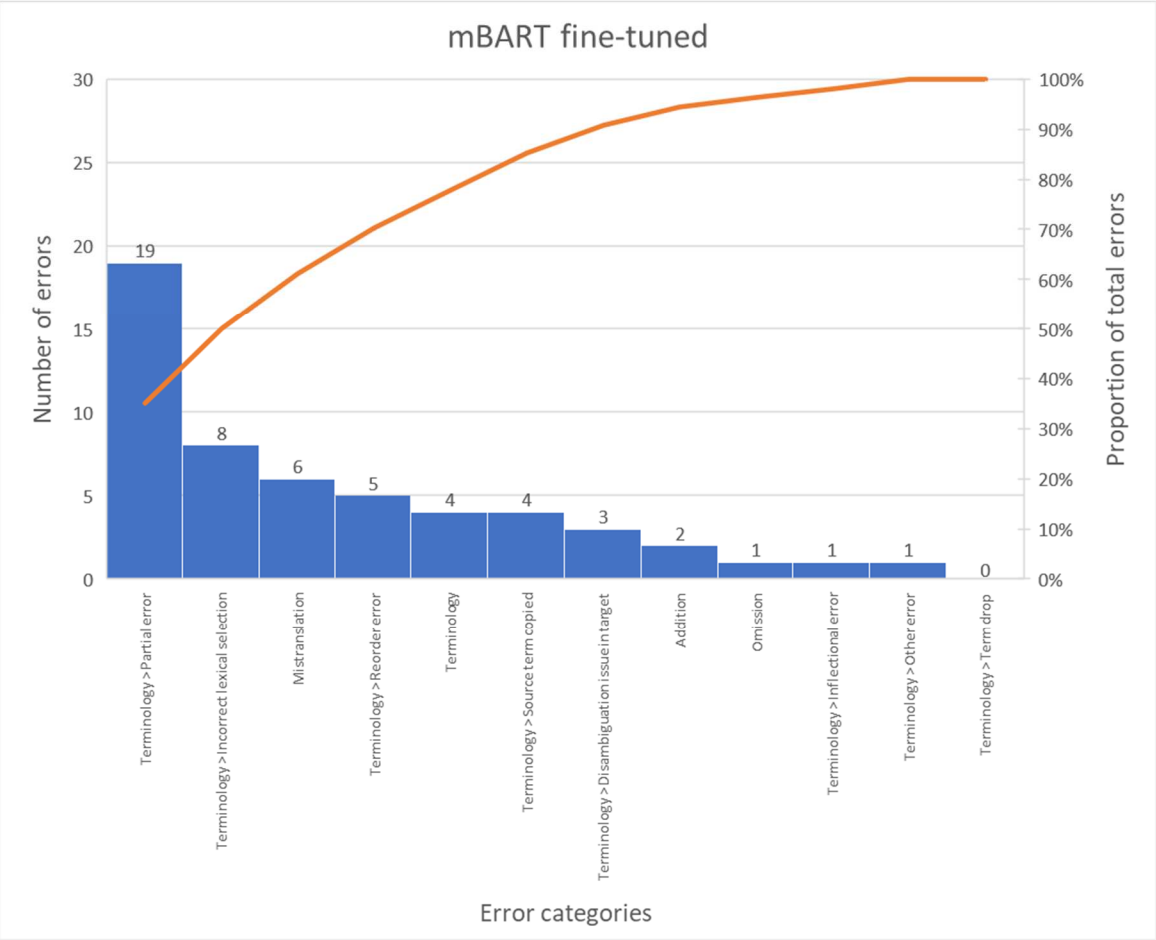

Figure 13: Accuracy errors in mBART fine-tuned output

The most frequent error category in the fine-tuned version of mBART is *Terminology > Partial error*, the same as in all outputs except for the reference translations. There are 19 errors in this category, making up 35% of the total amount of errors. An example of a *Major* error for this category is the following: “Teaching Institute of Public Health” in English was translated as “institut za obuku javnog zdravlja” in Croatian, whereas it should be “Nastavni zavod za javno zdravstvo”. While “Public Health” was translated correctly, the translation of “Teaching Institute” is incorrect.

The second frequent error category is *Terminology > Incorrect lexical selection* with eight errors. Together with the first category it makes up exactly 50% of the total amount of errors. An example of a *Major* error in the category *Terminology > Incorrect lexical selection* is the translation of the English term “lung lobes” into “plućni čvorovi” in Croatian, which is a different part of the lungs than the correct Croatian term “plućni režnjevi”.

The third frequent error category is *Mistranslation* with six errors. An example of a Major error is the translation of the English subtitle "Methods" into "Lijekovi" in Croatian (meaning "drugs" in English), whereas it should be "Metode". The most frequent three categories make up around 61% of all errors. There is only one error category without any errors.

4.2. Fluency errors

Fluency errors concern "the well-formedness of the target language, regardless of the transmission of content and meaning from the source into the target sentence." (Van Brussel et al., 2018, p. 3802) For this project a total of three error categories was considered when annotating *Fluency* errors in the reference translations as well as the outputs of the four NMT models. Figure 14 shows the findings of *Fluency* errors in the reference translations, where a total number of 42 *Fluency* errors was found.

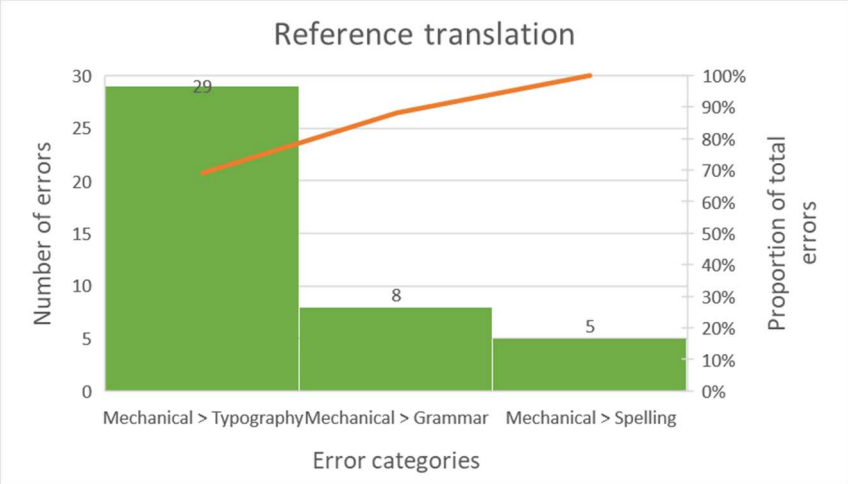

Figure 14: Fluency errors in reference translations

The most frequent error category in the reference translations is *Mechanical > Typography* with 29 errors. It has more than three times more errors than the second category (8 errors) and makes up around 69% of all errors. There are only *Minor* errors in the category *Mechanical > Typography*. An example of an error is the following: "n=31" in Croatian. There are no spaces before and after the equal sign "=".

The second frequent error category is *Mechanical > Grammar*, with eight errors. Again, there are only *Minor* errors in this category. An example of an error is the translation of "obese CAD patients" in English into "pretili bolesnici s IBS" in Croatian. Here, the suffix "-om" is missing after the abbreviation "IBS". The correct translation would be "pretili bolesnici s IBS-om".

The third frequent error category is *Mechanical* > *Spelling*, with five *Minor* errors. An example of an error in this category is the incorrect spelling of "mL" in Croatian. Its correct spelling would be "ml".

Figure 15 shows the findings of *Fluency* errors in the output of the NMT model Tilde, where a total number of 57 *Fluency* errors was found.

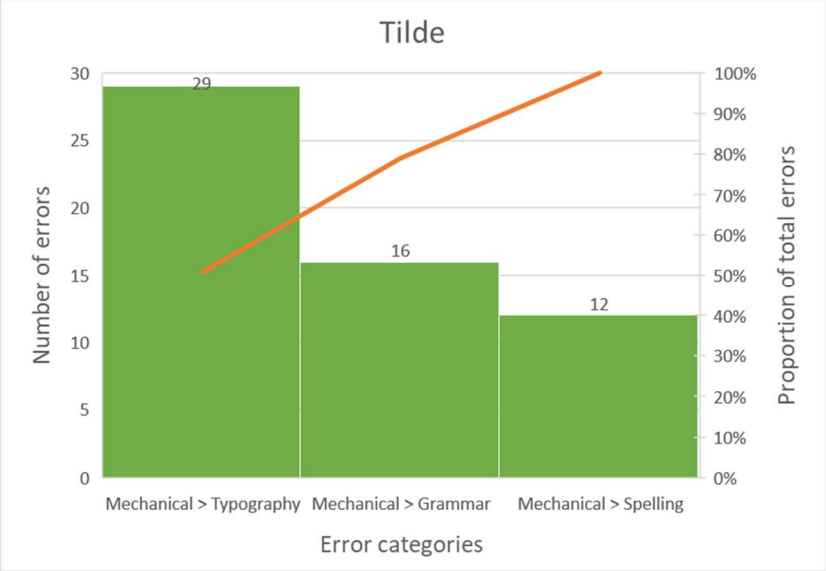

Figure 15: Fluency errors in Tilde output

The most frequent error category in the output of Tilde is *Mechanical* > *Typography* with 29 errors. Even though there is an equal number of errors in this category in the reference translations, here they make up only around 51% of all errors, compared to 69% in the reference translations. Again, there are only *Minor* errors in this category. An example of a *Mechanical* > *Typography* error is the following: In the Croatian output "siječanj-veljača" there are no spaces before and after the hyphen.

The second frequent error category is *Mechanical* > *Grammar*, with 16 *Minor* errors. An example of an error in this category is the translation of "subjects, […] mean age 59.43" in English into "ispitanici, […] prosječnom dobi 59,43" in Croatian. In this example, the declination is incorrect. The correct translation would be "ispitanici, […] prosječne dobi 59,43".

The third frequent error category is *Mechanical* > *Spelling*, with 12 *Minor* errors. An example of this category is the incorrect spelling of "irizin" in Croatian. The correct spelling would be "irisin".

Figure 16 shows the findings of *Fluency* errors in the output of the NMT model eTranslation, where a total number of 44 *Fluency* errors was found.

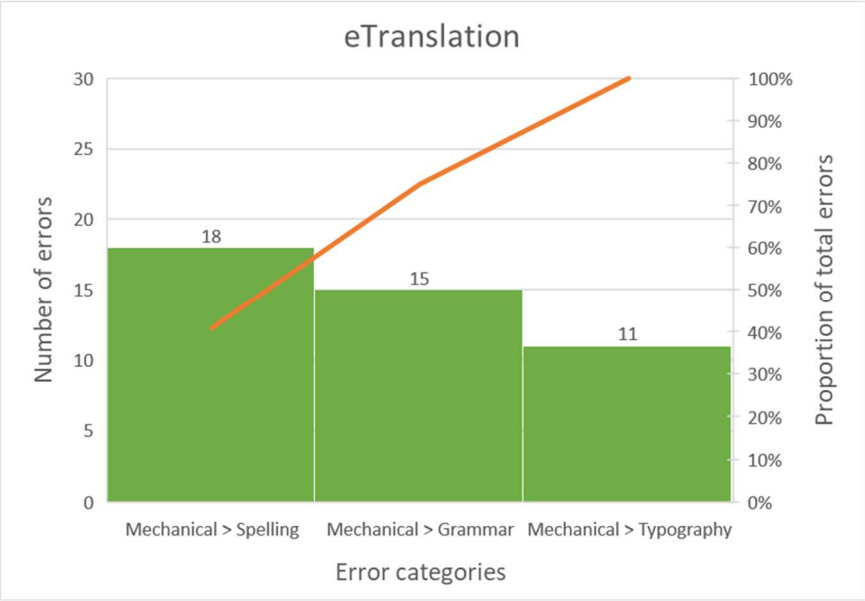

Figure 16: Fluency errors in eTranslation output

The most frequent error category in the output of eTranslation is *Mechanical > Spelling* with 18 errors, making up around 41% of the total amount of errors. An example of a *Minor* error in this category is the translation of "pneumothorax" in English into "pneumotooraks" in Croatian. However, the correct spelling of that word would be "pneumotoraks". An example of a *Major* error is the translation of "appendectomy" in English into "amendektomija" in Croatian, whereas the correct translation would be "apendektomija". The difference between these two examples of errors and their levels of severity is the result the incorrect spelling has on the understanding of the target output. While "pneumotooraks" has an extra "o" in the word, it can still be identified as the correct word "pneumotoraks". On the contrary, "amendektomija" has an "m" instead of an "p", changing the word in a way that the understanding is endangered.

The second frequent error category is *Mechanical > Grammar* with 15 errors. An example of a *Minor* error in this category is the following: "patients with microalbuminuria […] and those without it" in English was translated as "bolesnici s mikroalbuminurijom […] i oni bez **njega**" in Croatian. The pronoun "njega" is masculine and refers to "mikroalbuminurija", which is a feminine noun. Hence, the pronoun should be feminine as well. The correct translation would be "nje". An example of a *Major* error in the category *Mechanical > Grammar* is the translation of "obese patients with CAD" in English into "bolesnici s

CAD-om pretili" in Croatian. Here, the word order is incorrect and should be "pretili bolesnici s CAD-om". This error does not fall under the error category *Accuracy* > *Terminology* > *Reorder error* because the example does not include a term but a general part of a sentence.

The third frequent error category is *Mechanical* > *Typography* with 11 *Minor* errors. An example of this category is the translation of "group 1" in English into "skupina 1." in Croatian. Here, no full stop after the number is needed because "1." would refer to "first group" and not "group 1" like in the source text.

Figure 17 shows the findings of *Fluency* errors in the output of the NMT model mBART general, where a total number of 117 *Fluency* errors was found, more than in any of the other outputs or the reference translations.

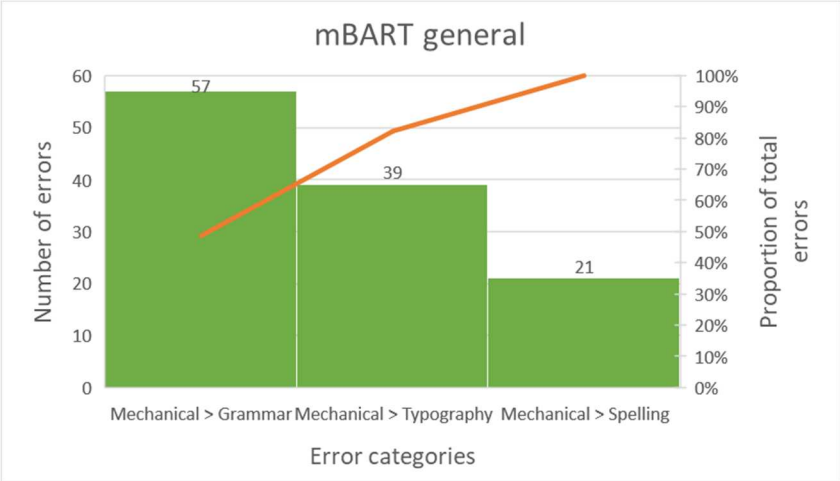

Figure 17: Fluency errors in mBART general output

The most frequent error category in the output of mBART general is *Mechanical* > *Grammar* with 57 errors, making up around 49% of the total amount of errors. There is only one *Major* error in this category, namely the following: "the urine albumin to creatinine ratio" in English was translated into "Ratio albumina urine na kreatininu" in Croatian. Besides other errors that do not fall under the category of *Mechanical* > *Grammar*, the part "na kreatininu" has an incorrect preposition and an incorrect word ending. The preposition "na" (meaning something like "on" in English) should be "i" (meaning "and" in English). Because of the new preposition, the word ending of "kreatininu" needs to change to "kreatinina". A *Minor* error in this category, which appeared several times, concerns the word order: "the aim of the study was" in English was translated into "cilj studije je bio" in Croatian, whereas in this case the correct word order would be "cilj studije bio je".

The second frequent error category is *Mechanical* > *Typography* with 39 *Minor* errors. An example of that category is the translation of the number "103.07" in the English source into "103.07" in the Croatian output. Decimal numbers require a comma in Croatian. Therefore, the correct translation of that number would be "103,07".

The third frequent error category is *Mechanical* > *Spelling* with 21 *Minor* errors. An example of that category is the incorrect spelling of the word "šestdeset" in Croatian. The correct spelling would be "šezdeset".

Figure 18 shows the findings of *Fluency* errors in the output of the NMT model mBART fine-tuned, where a total number of 79 *Fluency* errors was found.

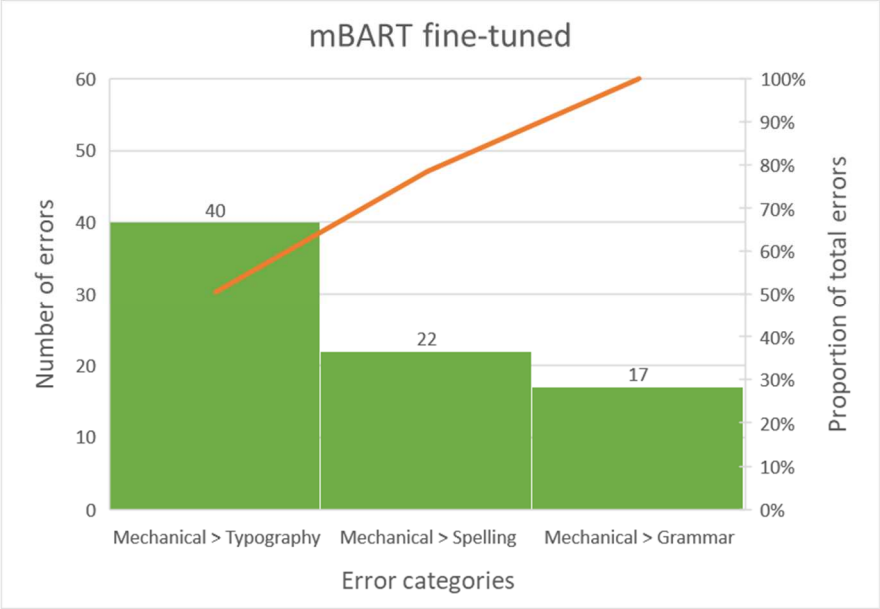

Figure 18: Fluency errors in mBART fine-tuned output

The most frequent error category in the output of mBART general is *Mechanical* > *Typography* with 40 *Minor* errors, making up around 51% of the total amount of errors. An example of that error category is the following: "95%" in Croatian. There needs to be a space between the number and percentage sign. The second frequent error category is *Mechanical* > *Spelling* with 22 *Minor* errors. One example is the incorrect spelling of "multivariabilne" in Croatian. The correct spelling would be "multivarijabilne". The third frequent error category is *Mechanical* > *Grammar* with 17 *Minor* errors. An example is the translation of "in various Zagreb departments" in English into "u različitim Zagrebskim odjelima" in Croatian. "Zagrebskim" is an incorrect grammatical form. The correct translation would be "u različitim Zagrebačkim odjelima".

4.3. Error severity

Alongside the different *Accuracy* and *Fluency* error categories, the severity of the errors was annotated in the reference translations as well as the NMT outputs. The error severity was based on the definitions by Lommel (2018):

- **Critical**: "[the errors] render a project unfit for purpose." (p. 120)
- **Major**: "[the errors] make the intended meaning of the text unclear in such a way that the intended user cannot recover the meaning from the text, but are unlikely to cause harm." (p. 120)
- **Minor**: "[the errors] do not impact usability." (p. 120-121)

Figure 19 shows the number of errors in the three severity levels (*Minor, Major* and *Critical*) for the reference translations and the NMT outputs.

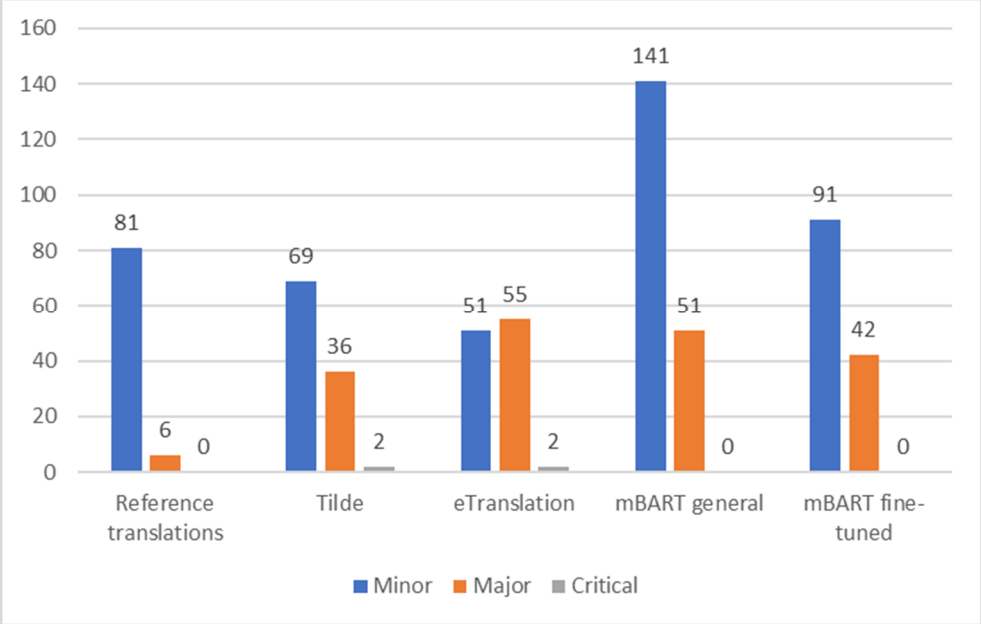

Figure 19: Error severity (Minor, Major and Critical) in reference translations and all NMT outputs

There are only four *Critical* errors in total, two in the output of Tilde and two in the output of eTranslation. Both *Critical* errors in the eTranslation output were assigned to errors containing wrong numbers. Especially in the medical domain this counts as a *Critical* error because a wrong number could make a project useless for its intended purpose or even cause danger when a certain procedure is carried out. The other two *Critical* errors in the output of Tilde are the translation of the verb "regress" in the English source with "pogoršati" in Croatian (meaning "deteriorate" in English) which has the opposite meaning of the source. The

second *Critical* error is the wrong translation of the unit of measurement "mg" in the English source into "min" in Croatian. Similar to the wrong translation of numbers this could cause confusion and have serious consequences for a medical procedure.

Most *Major* errors can be found in the output by eTranslation (55 errors), whereas the reference translations have the least number of *Major* errors (6 errors). The output of mBART general has clearly the most *Minor* errors (141 errors), while the output of eTranslation has the least (51 errors).

What is worth mentioning about the error severity in the reference translations is the low number of *Major* errors, namely only six. This could have been expected, considering that the reference translations are produced by humans and serve as a *golden standard* in this project. However, there are still 81 *Minor* errors in the reference translations. This calls into question their status as a *gold standard.* This will be discussed in more detail in section 5.3.

The output of Tilde has 69 *Minor* errors which is only a little less than the double of *Major* errors (36 errors), while the output of eTranslation has almost the same number of *Minor* (51 errors) and *Major* errors (55 errors). The output of mBART general has much more *Minor* errors than any other output, namely 141 errors, which is almost three times more than the number of *Major* errors (51 errors). This is because the output of mBART general has 117 *Fluency* errors (much more than the reference translations or any of the other NMT outputs) and these errors can usually be marked as *Minor*. For example, errors in the error category *Fluency > Mechanical > Typography* concerning spacing before and after punctuation marks like "=" or ">" are always marked as *Minor*. Finally, the output of mBART fine-tuned has almost the double number of *Minor* errors (91 errors) than *Major* errors (42 errors). The error severity can serve as a guidance for deciding which errors should be fixed first during the process of post-editing. Moreover, it will also play an essential role when deciding if certain error categories fall under light or full post-editing. This will be discussed in more detail in section 5.4.

4.4. Comparison

The errors in the reference translations and the NMT outputs were annotated according to the two categories *Accuracy* and *Fluency*. A total of 288 *Accuracy* errors was found. Figure 20 shows a comparison of the number of *Accuracy* errors in the reference translations and four NMT outputs divided into the error categories.

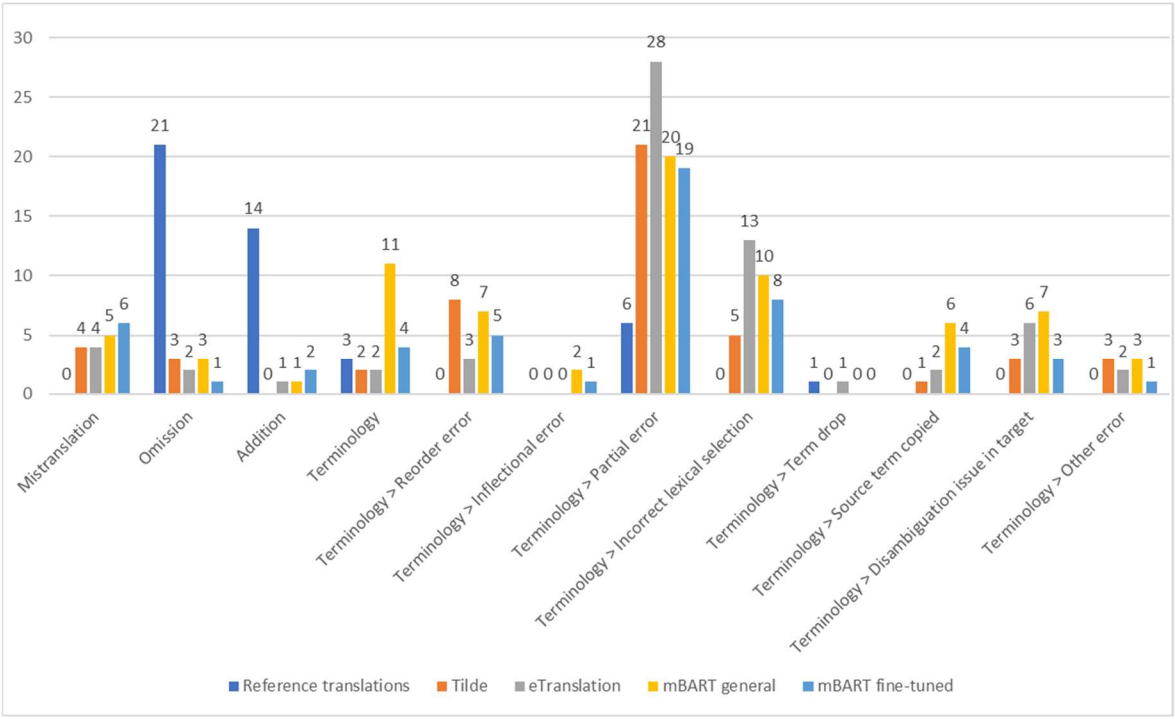

Figure 20: Comparison of Accuracy errors in reference translations and all NMT outputs

In the category *Mistranslation* there is a similar number of errors for all NMT outputs (4-6 errors). Only the reference translations include no error in this category. There are significantly more *Omission* errors in the reference translations (21 errors) than in any of the NMT outputs (1-3 errors). The same applies to *Addition* errors. While the output of Tilde has no errors in this category, the other NMT outputs have one or two errors. In the reference translations there are 14 *Addition* errors. The reason for the high number of errors in these two categories in the reference translations will be discussed in section 5.3. In the general error category *Terminology,* the output of mBART general has a significantly higher number of errors (11 errors) than the other outputs (2-4 errors). Concerning *Terminology > Reorder errors*, the outputs of Tilde and mBART general have slightly more (7-8 errors), while there are no errors in the reference translations. Errors in the category *Terminology > Inflectional error* can only be found in mBART general and mBART fine-tuned, but they are very few (1-2 errors). The category *Terminology > Partial error* is the most frequent error category

for all NMT outputs. Most errors in this category are found in the output of eTranslation (28 errors), the least in the reference translations (6 errors). In the *category Terminology > Incorrect lexical selection* the output of eTranslation has most errors (13 errors), while there are no errors in this category in the reference translations. Concerning the category *Terminology > Term drop* there is only one error in the reference translations and the output of eTranslation respectively. In the category *Terminology > Source term copied* there are only a few errors in the NMT outputs, most in the output of mBART general (6 errors) and none in the reference translations. The distribution is very similar in the category *Terminology > Disambiguation issue in target:* the output mBART general has most errors (7 errors) and the reference translations have none. Finally, there are only few errors (1-3 errors) in the category *Terminology > Other error* in the NMT outputs and none in the reference translations.

To sum up, by far the most *Accuracy* errors in the reference translations are found in the categories *Omission* (21 errors) and *Addition* (14 errors). The third frequent error category in the reference translations is *Terminology > Partial error*, which only has six errors. In contrast, most *Accuracy* errors in the NMT outputs are found in the error categories *Terminology > Partial error* and *Terminology > Incorrect lexical selection.* The least represented error categories are *Mistranslation*, *Terminology > Inflectional error*, *Terminology > Term drop* and *Terminology > Other error.*

Concerning the second main error category *Fluency,* a total of 339 errors was found in the reference translations and the NMT outputs. Figure 21Figure 20 shows a comparison of the number of *Fluency* errors in the reference translations and four NMT outputs divided into the error categories.

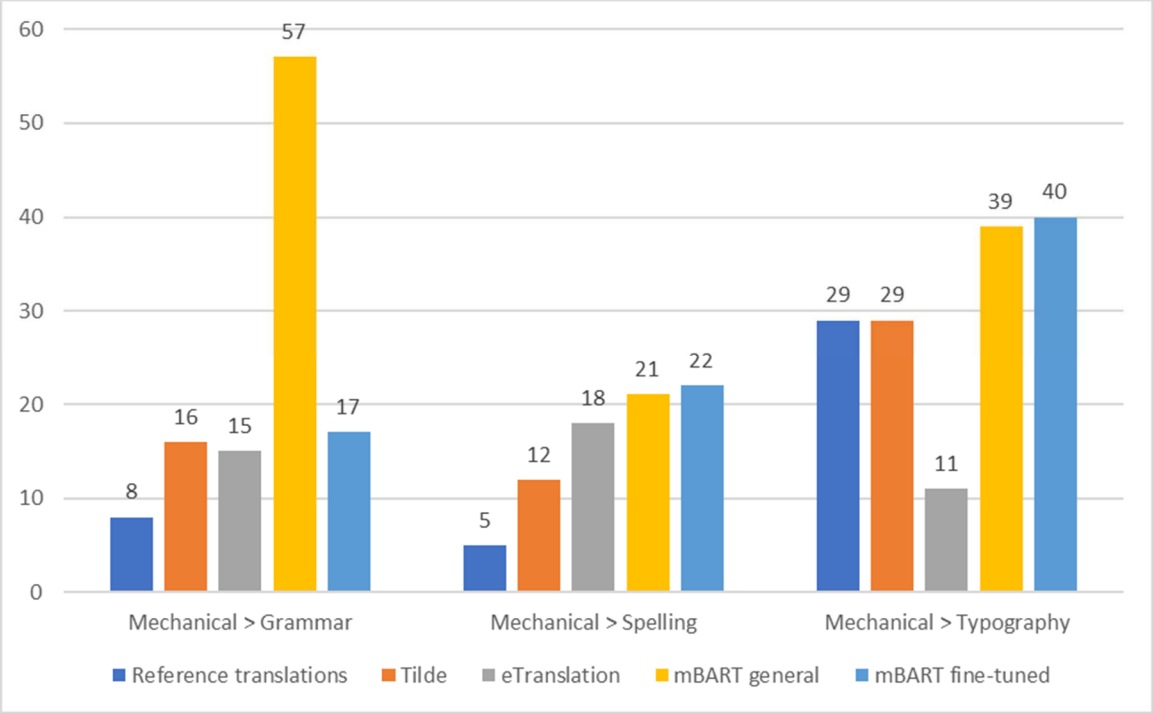

Figure 21: Comparison of Fluency errors in reference translations and all NMT outputs

In the category *Mechanical > Grammar* by far the most errors are found in the output of mBART general (57 errors). It has three times more errors than the other NMT outputs, where the number of errors is quite similar (15-17 errors). The least number of errors in this category can be found in the reference translations (8 errors). Concerning the category *Mechanical > Spelling* most errors are found in the outputs of mBART general (21 errors) and mBART fine-tuned (22 errors). The output of eTranslation has slightly less errors (18 errors) while the output of Tilde (12 errors) and the reference translations (5 errors) have the least number of errors. In the category *Mechanical > Typography* most errors are found in the outputs of mBART general (39 errors) and mBART fine-tuned (40 errors). The output of Tilde and the reference translations have the same number of errors (29 errors) while the output of eTranslation has by far the least number of errors (11 errors).

The low number of errors in the category *Mechanical > Typography* in the output of eTranslation stands out compared to the other NMT outputs and the reference translations in this category. However, the number of errors in the output of eTranslation is similarly distributed among the three error categories (15, 18 and 11 errors). In the reference translations the distribution of errors is unequal: While there are only eight and five errors in two categories, the category *Mechanical > Typography* has significantly more errors (29 errors). In the output of Tilde, the errors are quite similarly distributed in the first two categories (16 and 12 errors) while there are twice as many errors in the third category (29 errors). Similar to the reference translations, the distribution of errors in the output of mBART general is unequal: While there are a lot of errors in the first category (57 errors), there are much less errors in the second (21 errors) but again more (39 errors) in the third category. The situation is comparable to the output of mBART fine-tuned: There is a similar number of errors in the first two categories (17 and 22 errors), while there are by far the most errors in the third category (40 errors).

5. Discussion

5.1. Tilde vs. eTranslation

Even though the NMT models Tilde and eTranslation were developed for different purposes, they are offered side by side as tools for the EU Council Presidency Translator[17]. While Tilde develops its own general, custom or adaptive MT systems according to different needs, eTranslation is a NMT model developed by and for the needs of translators at the EU (Tilde Senior Account Manager, personal communication, October 27, 2022). The results of the error annotation in the outputs of Tilde and eTranslation will be analysed in the following paragraphs.

Concerning the *Accuracy* errors, the outputs of the NMT models show the following results: While in the output of Tilde 50 *Accuracy* errors are found in total, the output of eTranslation contains 64 *Accuracy* errors (see Figures 22 and 23). So, the difference between the total amount of errors is 14.

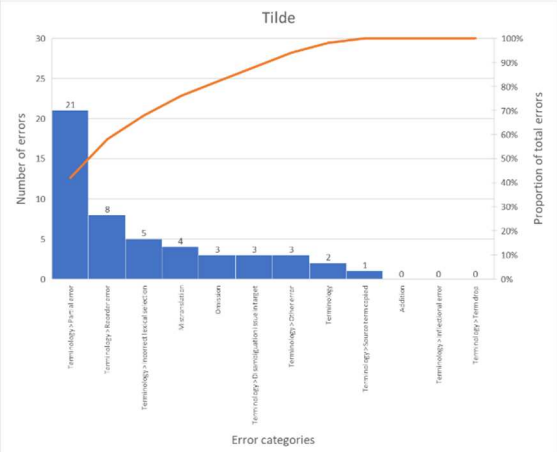

Figure 22: *Accuracy* errors in the output of Tilde

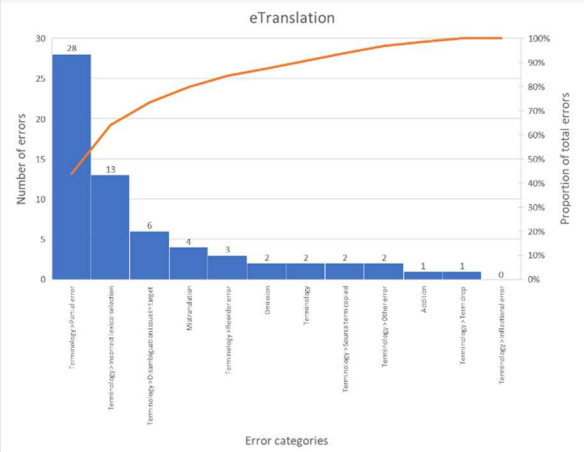

Figure 23: *Accuracy* errors in the output of eTranslation

In both outputs the most frequent error category is *Terminology > Partial error.* While the output of Tilde has 21 errors in this category, the output of eTranslation shows 28 errors. However, in both outputs the first category makes up a similar proportion of total errors: in Tilde's output the first error category makes up 42% of all errors and in eTranslation's output 44%. Compared to the second frequent error category, Tilde's output has almost three times

[17] See https://www.tilde.com/products-and-services/machine-translation/de-presidency, retrieved 8 November 2022.

the number of errors in the first category: 21 *Terminology > Partial errors* and eight *Terminology > Reorder errors*. In contrast, eTranslation's output has less than the double number of errors in the first category compared to the second category: 28 *Terminology > Partial errors* and 13 *Terminology > Incorrect lexical selection errors.*

The second frequent error category differs in both outputs: as stated above, Tilde's second frequent category is *Terminology > Reorder error* (8 errors) and eTranslation's second frequent category *Terminology > Incorrect lexical selection error* (13 errors). The first two categories make up 56% of all errors in the output of Tilde, while they make up 64% of all errors in the output of eTranslation. This means that by post-editing the first two error categories in the output of eTranslation, almost two thirds of all errors can be corrected whereas in the output of Tilde only a little more than half of the total amount of errors would be fixed.

The third frequent error category in the output of Tilde is *Terminology > Incorrect lexical selection error* with five errors, which together with the first two categories makes up 68% of all errors. In the output of eTranslation the third frequent error category is *Terminology > Disambiguation issue in target* with six errors, which makes up around 73% of all errors. This means that by editing the errors in the first three categories in the output of eTranslation more errors are fixed than when the first three categories in the output of Tilde are fixed.

It is visible that most errors in both outputs concern terminology. The error categories *Terminology > Partial error, Terminology > Reorder error, Terminology > Incorrect lexical selection error* and *Terminology > Disambiguation issue in target* are most represented. In the outputs of both NMT models the fourth frequent category (and the first that does not concern Terminology) is *Mistranslation* with four errors in each output. There are only three errors in the category *Omission* in the output of Tilde while there are only two in the output of eTranslation. There is no *Addition* error in Tilde's output and only one in eTranslation's output. This is the biggest difference compared to the reference translations which will be discussed in section 5.3.

To sum up, eTranslation produces more *Accuracy* errors than Tilde. However, by editing the first three error categories, 73% of all errors can be fixed in the output of eTranslation, while slightly less (68% of all errors) can be fixed in the output of Tilde. Apart from this both NMT models produce similar results when it comes to *Accuracy* errors with most errors concerning terminology.

Regarding the *Fluency* errors, the outputs of the NMT models Tilde and eTranslation show the following results: while in the output of Tilde 57 *Fluency* errors are found in total, the output of eTranslation contains 44 *Fluency* errors (see Figures 24 and 25Figure 23). So, the difference between the total amount of errors is 13.

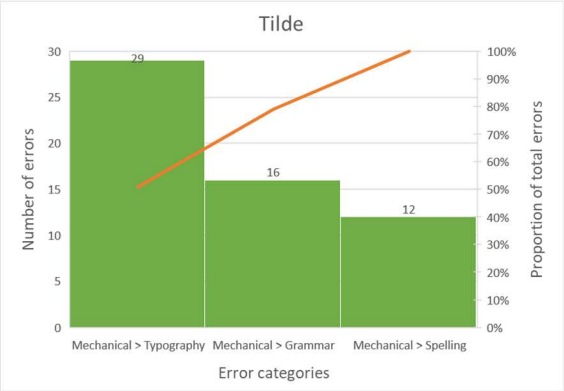

Figure 24: *Fluency* errors in the output of Tilde

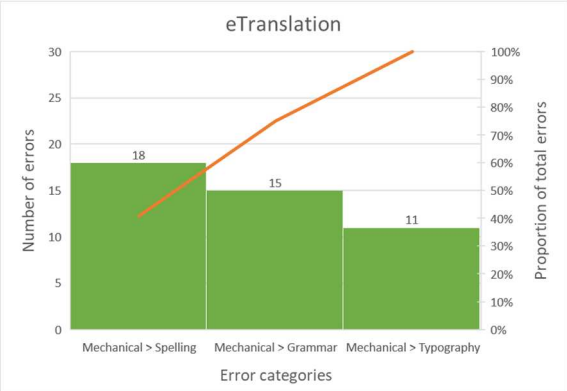

Figure 25: *Fluency* errors in the output of eTranslation

In the output of Tilde, the most frequent error category is *Mechanical > Typography* with 29 errors while in the output of eTranslation it is *Mechanical > Spelling* with 18 errors. This first error category makes up 51% of all errors in Tilde's output and 41% in eTranslation's output. This means that by editing the errors in the first category, half of all errors can be fixed in Tilde's output and less than half in eTranslation's output.

The second frequent error category in both outputs is *Mechanical > Grammar*, with 16 errors in Tilde's output and 15 errors in eTranslation's output. Even though the number of errors is almost the same in the second category for both outputs, their relation to the number of errors in the first category differs: while there are almost twice as many errors in the first category in Tilde's output, there are only three errors more in the first category in eTranslation's output. The first two categories make up 79% of the total amount of errors in the output of Tilde and 75% in the output of eTranslation. This means that by editing the errors in the first two categories more errors can be corrected in Tilde's output.

Finally, the third frequent error category is *Mechanical > Spelling* with 12 errors in the output of Tilde and *Mechanical > Typography* with 11 errors in the output of eTranslation. The third category has a similar number of errors for both outputs.

To sum up, Tilde produces more *Fluency* errors than eTranslation. However, by editing the first error category half of all errors (51%) can be fixed in the output of Tilde while only 41% can be fixed in the output of eTranslation. This is because there are more errors in the first category in Tilde's output in relation with the total amount of errors. Also, errors in

the category *Mechanical* > *Typography* might be easier to correct than errors in the category *Mechanical* > *Spelling* because typography errors often concern spacing before and/or after punctuation while spelling errors might require the post-editor to research for the right spelling of a word.

As Table 2 shows, the outputs of Tilde and eTranslation have some differences concerning the error severity.

Table 2: Comparison of the error severity in the outputs of Tilde and eTranslation

Error severity	**Tilde**	**eTranslation**
Minor	69	51
Major	36	55
Critical	2	2
Total	**107**	**108**

While both outputs have a similar number of total errors (107 errors in Tilde's output and 108 errors in eTranslation's output), their distribution among the three levels of error severity varies. The output of Tilde has 69 *Minor* errors, whereas the output of eTranslation only has 51. This is mainly because Tilde's output has more *Fluency* errors (57 errors) than eTranslation's output (44 errors) and they tend to be *Minor* errors rather than *Major*. Concerning the *Major* errors, the situation is reversed: Tilde's output shows 36 errors while eTranslation has 55 errors. Again, this is due to the higher number of *Accuracy* errors in the output of eTranslation (64 errors) compared to the output of Tilde (50 errors), which are more often categories as *Major* errors than *Fluency* errors. Moreover, the output of eTranslation has only slightly more *Major* errors (55 errors) than *Minor* errors (51 errors) while the output of Tilde has almost twice as many *Minor* errors (69 errors) than *Major* errors (36 error). Both outputs include two *Critical* errors respectively.

To sum up, in this project eTranslation produced more *Accuracy* errors than Tilde, while Tilde produced more *Fluency* errors than eTranslation. Therefore, eTranslation led to more *Major* errors and Tilde to more *Minor* errors, while both included two *Critical* errors. Moreover, eTranslation produced only slightly more *Major* errors than *Minor* errors, while Tilde had significantly more *Minor* errors than *Major* errors. In total, both NMT models produced almost the same number of errors. Which of the errors require light and which need full post-editing will be discussed in section 0.

5.2. mBART general vs. mBART fine-tuned

The NMT model mBART general is a "multilingual encoder-decoder (sequence-to-sequence) model" (von Platen, n.d.) and was used for this project alongside a fine-tuned version for the medical domain. The fine-tuned version was trained with a multilingual corpus (English-Croatian) made out of PDF documents from the European Medicines Agency (EMEA) (Prokopidis & Papavassiliou, 2020).

Concerning the *Accuracy* errors the outputs of the NMT models mBART and mBART fine-tuned show the following results: While in the output of mBART general 75 *Accuracy* errors are found in total, the output of mBART fine-tuned contains 54 *Accuracy* errors (see Figures 26 and 27). So, the difference between the total amount of errors is 21.

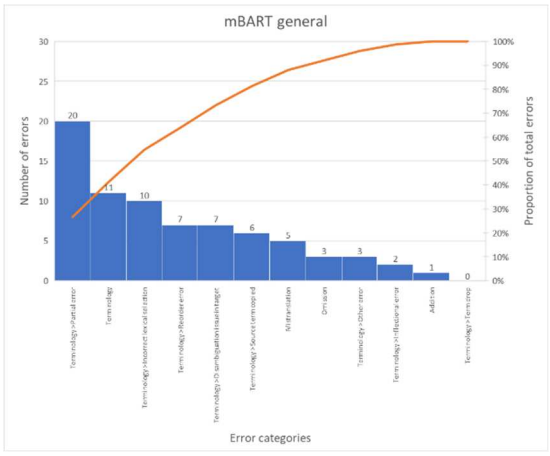

Figure 26: *Accuracy* errors the output of mBART general

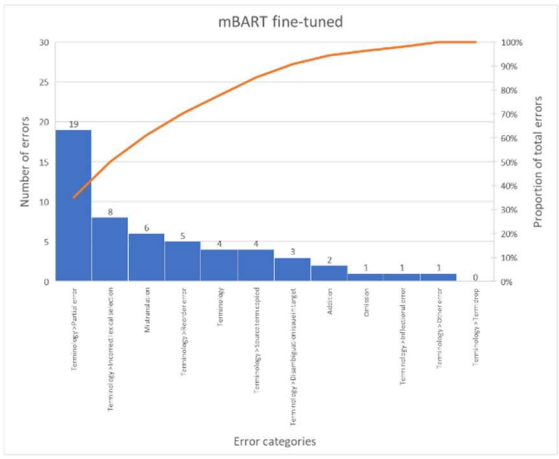

Figure 27: *Accuracy* errors in the output of mBART fine-tuned

In both outputs the most frequent error category is *Terminology > Partial error* with 20 errors in the output of mBART general and 19 errors in the output of mBART fine-tuned. However, while in mBART general's output the first error category makes up only 27% of the total amount of errors (less than one third), in the output of mBART fine-tuned it makes up 35% of all errors (more than one third). This has to do with the higher total number of errors in the output of mBART general.

The second frequent error category in the output of mBART general is *Terminology* with 11 errors. Together with the first error category it makes up 48% of all errors. In the output of mBART fine-tuned the second frequent category is *Terminology > Incorrect lexical selection* with eight errors. Together with the first category it makes up exactly 50% of all errors. Here, both outputs show similar results: By editing the first two categories, half of all errors can be corrected.

The third frequent error category in mBART general's output is *Terminology > Incorrect lexical selection* with ten errors which is only one error less than in the second category. The first three error categories together make up around 55% of all errors. The third frequent category in the output of mBART fine-tuned is *Mistranslation,* the first category not directly concerned with terminology. There are six errors in this category, making up around 61% together with the first two categories. This means that by editing the first three categories, more errors are fixed in the output of mBART fine-tuned than in the output of mBART general.

It is visible that most errors in both outputs concern terminology. The error categories *Terminology > Partial error, Terminology* and *Terminology > Incorrect lexical selection error* are most represented, with the category *Mistranslation* following shortly after. The error categories *Omission* and *Addition* are not so dominant: there are only three *Omission* errors in mBART general's output and one *Omission* error in the output of mBART fine-tuned. Similarly, there is only one *Addition* error in the output of mBART general and two *Addition* errors in the output of mBART fine-tuned. There are no errors in the category *Terminology > Term drop* in both outputs and only a low number of errors in the remaining categories (1-5 errors per category in the output of mBART general and 1-7 errors in the output of mBART fine-tuned).

To sum up, mBART general produces more *Accuracy* errors than mBART fine-tuned. This shows that for this project fine-tuning the NMT model mBART led to fewer *Accuracy* errors in total. Also, by editing the first three categories more errors in total can be fixed in the output of mBART fine-tuned (61% in contrast to 55% in the output of mBART general). Therefore, mBART fine-tuned surpasses mBART general in two aspects: the total amount of *Accuracy* errors is lower, and more errors can be fixed by editing the first three categories. Apart from this both NMT models produce most errors concerning terminology.

Regarding the *Fluency* errors the outputs of the NMT models mBART general and mBART fine-tuned show the following results: While in the output of mBART general 117 *Fluency* errors are found in total, the output of mBART fine-tuned contains 79 *Fluency* errors (see Figure 24 Figures 28 and 29Figure 25Figure 23). So, the difference between the total amount of errors is 38.

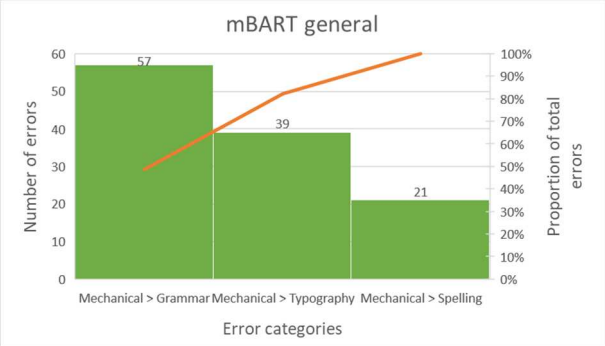

Figure 28: *Fluency* errors the output of mBART general

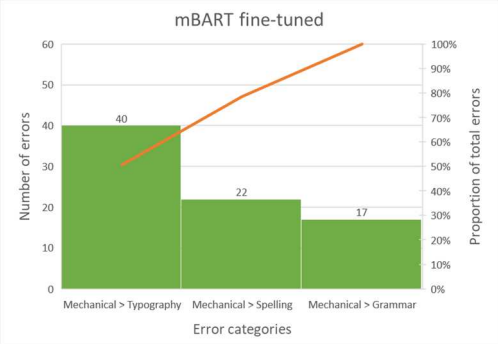

Figure 29: *Fluency* errors the output of mBART general

In the output of mBART general, the most frequent error category is *Mechanical > Grammar* with 57 errors while in the output of mBART fine-tuned it is *Mechanical > Typography* with 40 errors. This first error category makes up 49% of all errors in mBART general's output and 51% in the output of mBART fine-tuned. This means that by editing the errors in the first category, in both outputs almost half of the errors can be fixed. However, errors in the category *Mechanical > Typography* might be easier to correct than errors in the category *Mechanical > Grammar* because typography errors often concern spacing before and/or after punctuation while grammatical errors might require the post-editor to research for the right grammatical form.

The second frequent category in the output of mBART general is *Mechanical > Typography* with 39 errors and *Mechanical > Spelling* with 22 errors in the output of mBART fine-tuned. There are not only almost twice as many errors in mBART general's output in this category compared to the output of mBART fine-tuned, but the relation of the number of errors in the second category compared to the first category differs: While the first category only has around one third more errors than the second in the output of mBART general there are almost twice as many errors in the first category compared to the second in the output of mBART fine-tuned. Nevertheless, the two categories together account for a similar share of the total errors: They make up 82% in the output of mBART general and 78% in the output of mBART fine-tuned. The third and last category is *Mechanical > Spelling* with

21 errors in mBART general's output and *Mechanical* > *Grammar* with 17 errors in the output of mBART fine-tuned.

To sum up, mBART general produces far more *Fluency* errors than mBART fine-tuned. By editing the errors in the first or in the first two categories a similar amount of errors can be fixed in both outputs. Moreover, mBART fine-tuned produces most errors in the category *Mechanical* > *Typography*, which can be edited more easily as was stated above. This means that mBART fine-tuned outperforms mBART general concerning *Fluency* errors as well.

As Table 3 shows, the outputs of mBART general and mBART fine-tuned have some differences concerning the error severity.

Table 3: Comparison of the error severity in the outputs of mBART general and mBART fine-tuned

Error severity	**mBART general**	**mBART fine-tuned**
Minor	141	91
Major	51	42
Critical	0	0
Total	**192**	**133**

The total number of errors varies between the two outputs: mBART general shows 192 *Accuracy* and *Fluency* errors, while mBART fine-tuned has 133 errors. The output of mBART general has both more *Minor* and *Major* errors than the output of mBART fine-tuned: mBART general includes 141 *Minor* errors while mBART fine-tuned only has 91. Again, this has to do with the higher number of *Fluency* errors in the output of mBART general (117 errors compared to 79 in the output of mBART fine-tuned), which tend to be categorised as *Minor* errors. mBART general also has slightly more *Major* errors (51 errors) compared to mBART fine-tuned (42). Both NMT models have more *Minor* errors than *Major* errors in their output. Neither of the outputs have any *Critical* errors.

To sum up, in this project mBART general includes both more *Accuracy* and more *Fluency* errors than mBART fine-tuned. mBART general also has both more *Minor* and more *Major* errors than mBART fine-tuned. Moreover, mBART general shows more errors in total. Also, by editing the first three categories in the output of mBART fine-tuned more errors in total can be fixed. Finally, the most frequent *Fluency* error category in the output

of mBART fine-tuned is *Mechanical* > *Typography*, a category that can be edited more easily than others. This means that mBART fine-tuned outperforms mBART general in all aspects. Which of the errors require light and which need full post-editing will be discussed in section 0.

5.3. Reference translations

When choosing the dataset for this research project one condition was that the English source texts have Croatian reference translations that can be used as a *gold standard*. When annotating the errors in the NMT outputs, the reference translations served as a basis for annotation decisions. However, their status as a *gold standard* needs to be questioned. As was mentioned at the beginning of this paper (see section 2.5), there is evidence that human translations do not always offer 100% quality, even after revision (Ciobanu et al., 2019; Daems & Macken, 2020). This is visible in the findings of this project, as well, and will be discussed in more detail in the next paragraphs. In total, 45 *Accuracy* errors are found in the reference translations (see Figure 30).

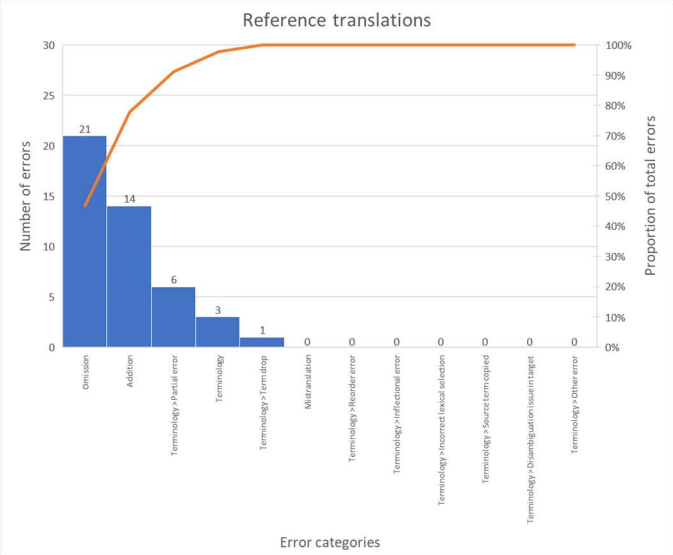

Figure 30: *Accuracy* errors in the reference translations

The most frequent error category in the reference translations is *Omission* with 21 errors and the second frequent category is *Addition* with 14 errors. One explanation for the high number of errors in these categories is the fact that human translators do not translate word-by-word, which is a more common strategy for machine translation. They rather add explanations when they are needed for better understanding or omit information that might seem irrelevant due to the context. They also often merge two segments into one. This is not an error per se,

but it was annotated as an error for this project. This is because any deviation from the source text was annotated as an error for better comparison with the other outputs. It can be deduced from this that omissions and additions do not mean that the reference translations are incorrect, but these types of errors do endanger the status of reference translations as a *gold standard* for annotating errors in NMT outputs or for generating automatic scores for reference-based MT evaluation algorithms.

Another aspect concerning omissions and additions is the question whether post-editing could be done monolingually, as well. Omissions and additions cannot be spotted monolingually but need the source text as a reference. Most errors in the categories *Omission* and *Addition* were found in the reference translations, which do not require post-editing but rather revision. In contrast, the NMT outputs showed only a low number of omissions and additions (see Table 4).

Table 4: *Omission* and *Addition* errors in the NMT outputs

NMT model	**Omission errors**	**Addition errors**
Tilde	3	0
eTranslation	2	1
mBART general	3	1
mBART fine-tuned	1	2

Considering the low number of errors in these two categories in the NMT outputs, monolingual post-editing might seem like a possibility. However, even just a few omissions and additions that get overlooked can have a great impact on the understanding of the target text. This is especially problematic with dense texts like medical abstracts that need to be precise and include all the information. Therefore, it is not advisable to do monolingual post-editing of medical abstracts for the language pair English-Croatian.

Coming back to the reference translations, no mistranslations are found which speaks for their status as a *gold standard.* However, still a few terminology errors are found (10 errors). Considering that these ten errors are distributed among 68 segments, this does not seem like a high number. Nevertheless, this shows that reference translations produced by human translators do not offer 100% quality.

This is also visible in the *Fluency* errors found in the reference translations. In total, 42 *Fluency* errors are found in the reference translations (see Figure 31).

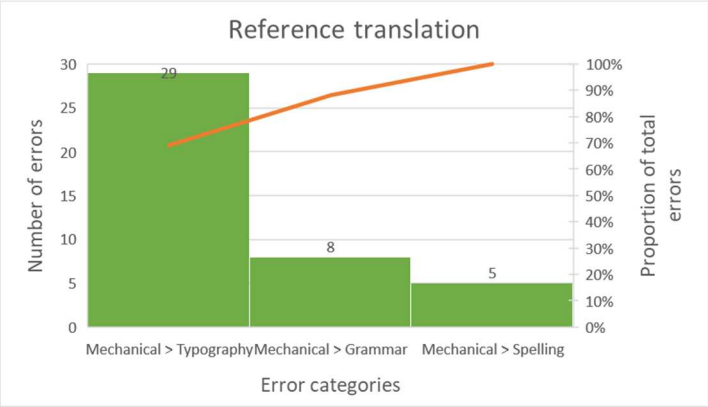

Figure 31: *Fluency* errors in the reference translations

The most frequent error category in the reference translations is *Mechanical > Typography* with 29 errors. Almost all of these errors concern the spacing before and after certain punctuation like for example “%”, “-“ or “<”. By editing this first category 69% of all errors can be corrected. The second frequent error category is *Mechanical > Grammar* with eight errors and the last category is *Mechanical > Spelling* with five errors. Even though all 42 *Fluency* errors are *Minor* errors and therefore “do not impact usability” (Lommel 2018, p. 120-121), they still make clear that the reference translations do not provide 100% quality.

As Table 5 shows, there are 87 errors in total in the reference translations whereof 81 are *Minor* errors and six are *Major* errors. There are no *Critical* errors in the reference translations. Even though there are only six *Major* errors that "make the intended meaning of the text unclear" (Lommel 2018, p. 120), they still need to be corrected to ensure a high quality of the translations. The same is true for *Minor* errors: even though a translation can be used with *Minor* errors for some purposes, other purposes require a higher quality and a translation without any errors. Which of these errors require light and which need full post-editing will be discussed in the next section (even though a traditional view would be that the reference translations need revision, not post-editing).

Table 5: Error severity in the reference translations

Error severity	**Reference translations**
Minor	81
Major	6
Critical	0
Total	**87**

5.4. Light vs. full post-editing

In the previous chapters the findings were presented for every single output, as well as in comparison with each other. The errors were assigned to the different error categories and severity levels. Now, the question arises if editing those errors would fall under the process of *light* or of *full* post-editing. As mentioned in section 2.5., light and full post-editing describe the final quality of the translation rather than the post-editing process. Nevertheless, there are some error types that usually fall under either light or full post-editing. Based on the post-editing guidelines summarised for this research in Table 1 in section 3.4., the different error categories represented in this project can be assigned to either light or full post-editing.

Table 6 shows the division of the error categories found in this research project into light post-editing on the left and full post-editing on the right side.

Table 6: First version of a division of error categories into light or full post-editing

Light post-editing	**Full post-editing**
Accuracy > Addition	Fluency > Mechanical > Grammar
Accuracy > Omission	Fluency > Mechanical > Typography
Accuracy > Mistranslation	Accuracy > Terminology > all categories
Fluency > Mechanical > Spelling	
Critical errors	

The four error categories *Accuracy > Addition, Accuracy > Omission, Accuracy > Mistranslation* and *Fluency > Mechanical > Spelling* can be assigned to light post-editing alongside errors in any error category with the severity level *Critical.* The first two error categories *Accuracy > Addition* and *Accuracy > Omission* correspond to the aspect of *Information* in the post-editing guidelines (see section 2.5.), which says the following for light post-editing: "Ensure that no information has been accidentally added or omitted" (TAUS 2016). This means that all additions and omission should be edited to ensure the translation quality that is needed after light post-editing.

The third error category *Accuracy > Mistranslation,* as well as any errors with the error severity *Critical* correspond to the aspect of *Accuracy* in the post-editing guidelines, according to which light post-editing should lead to the following two points in the final translation: "TT communicates the same meaning as ST" (TAUS 2016) and "Factually accurate" (Densmer 2014). Therefore, mistranslations should be edited in light post-editing as well to ensure that the same meaning is communicated and that all facts are correct.

Finally, the fourth error category *Fluency > Mechanical > Spelling* should be edited in light post-editing, as well, considering the aspect *Spelling* in the post-editing guidelines: "Basic rules regarding spelling apply" (TAUS 2016).

The other three error categories *Fluency > Mechanical > Grammar, Fluency > Mechanical > Typography* and *Accuracy > Terminology > all categories* can be assigned to full post-editing. The first error category *Fluency > Mechanical > Grammar* corresponds to the aspect of *Grammar* in the post-editing guidelines, which for light post-editing "[m]ay not be perfect" (TAUS 2016) whereas for full post-editing it should "[a]im for grammatically […] correct translation" (TAUS 2016). This means that any grammatical errors should be corrected in full post-editing only.

The second error category *Fluency > Mechanical > Typography* should be edited in full post-editing as well considering the aspect *Punctuation* in the post-editing guidelines: "Basic rules regarding […] punctuation and hyphenation apply" (TAUS 2016).

The error category *Accuracy > Terminology > all categories* includes all error categories concerning terminological errors. In total this concerns the following nine error categories:

- Accuracy > Terminology
- Accuracy > Terminology > Reorder error
- Accuracy > Terminology > Inflectional error
- Accuracy > Terminology > Partial error
- Accuracy > Terminology > Incorrect lexical selection
- Accuracy > Terminology > Term drop
- Accuracy > Terminology > Source term copied
- Accuracy > Terminology > Disambiguation issue in target
- Accuracy > Terminology > Other error

These nine error categories would fall under full post-editing according to the aspects *Accuracy* and *Terminology* in the post-editing guidelines. Concerning *Accuracy*, the full post-edited translation should fulfil the following goals: "TT communicates the same meaning as ST" (TAUS 2016) and "Absolutely accurate" (Densmer 2014). In the post-editing guidelines, *Terminology* plays the following role for light post-editing: "No need to spend too much time researching if incorrect" (Mesa-Lao 2013) and "Be consistent" (Densmer 2014). On the contrary, for full post-editing *Terminology* should be dealt with in the following way: "Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of 'Do Not Translate' terms" (TAUS 2016).

Looking very closely at these guidelines for terminology errors, the problem arises that they remain very vague. Especially the phrasing of the light post-editing guideline "[n]o need to spend too much time researching if incorrect" (Mesa-Lao 2013) leaves the question open what "too much time" means and if terminology should be researched at all. Correctly translated terminology is not specifically required in light post-editing but only in full post-editing. However, a "[f]actually accurate" (Densmer 2014) translation is a requirement of light post-editing. Now, it can be argued that incorrect or partly incorrect terminology (which was the most frequent error category for all outputs except the reference translation) endangers "[f]actually accurate" translations. This is especially true for terminology errors with the severity *Major*. While *Minor* terminology errors might be ignored in light post-editing and would not change the meaning of the source text, *Major* terminology errors could indeed change the meaning, transfer inaccurate information, and consequently lead to confusion and misunderstanding.

Therefore, the previously presented division of the error categories into light and full post-editing in Table 6 (see p. 74) should be revised. *Major* terminology errors should be included in the process of light post-editing while *Minor* terminology errors can be corrected in full post-editing only. The revised final version of the division into light and full post-editing is presented in Table 7. Errors in the error category *Accuracy > Terminology > all categories* with the error severity *Major* are placed under light post-editing while errors in the same categories with the error severity *Minor* are placed under full post-editing.

Table 7: Final version of a division of error categories into light or full post-editing

Light post-editing	**Full post-editing**
Accuracy > Addition	Fluency > Mechanical > Grammar
Accuracy > Omission	Fluency > Mechanical > Typography
Accuracy > Terminology > all categories (Major errors)	Accuracy > Terminology > all categories (Minor errors)
Accuracy > Mistranslation	
Fluency > Mechanical > Spelling	
Critical errors	

Based on the division into the two categories of light and full post-editing in Table 7, the errors found in the reference translations and all NMT outputs are added up and presented as absolute numbers, as well as the percentage share of the total number of errors in Table 8. It is important to keep in mind that in full PE all errors need to be corrected, including the errors falling under the category of light PE. In contrast, the absolute number and percentage of errors in full PE represent the errors that are corrected in full PE only (and not in light PE). However, a full PE procedure would include correcting 100% of the errors (if the client requests it).

Table 8: Number and percentage share of errors in the categories light and full post-editing for all outputs

Output	**Light post-editing**	**Full post-editing**
Reference translations	46 errors (~53.9%)	41 errors (~47.1%)
Tilde	53 errors (~49.5%)	54 errors (~50.5%)
eTranslation	67 errors (~62.0%)	41 errors (~38.0%)
mBART general	74 errors (~38.5%)	118 errors (~61.5%)
mBART fine-tuned	65 errors (~48.9%)	68 errors (~51.1%)

The output of eTranslation has most errors falling under the category of light post-editing, namely 67 out of 108 (~62.0% of all errors in this output). In contrast, the output of mBART general has the least number of errors falling under the category of light post-editing: 74 errors out of 192 (~38.5% of all errors in this output).

Figure 32 shows a visual comparison of the percentage share of errors in the outputs requiring light or full post-editing.

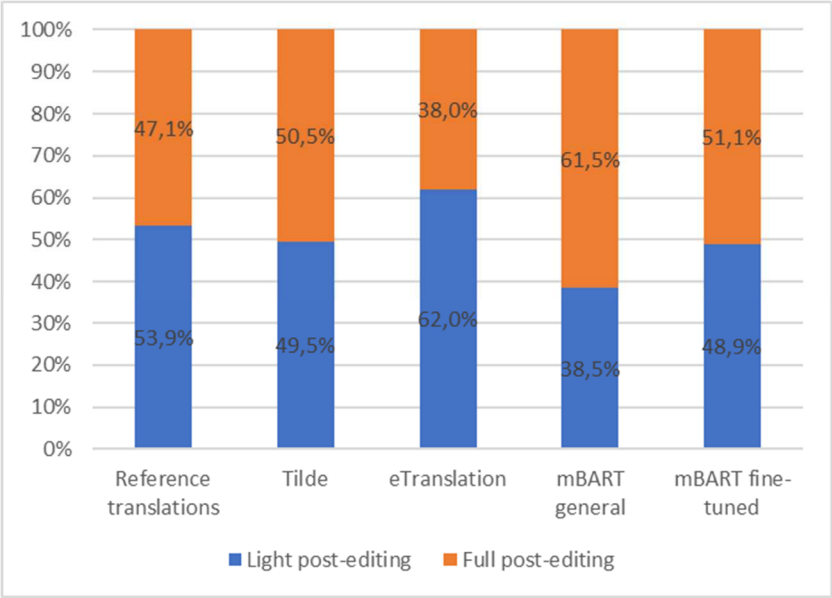

Figure 32: Percentage share of errors in the categories light and full post-editing for all outputs

As mentioned above, the output of eTranslation (~62.0% of errors requiring light post-editing) and mBART general (~38.5% of errors requiring light post-editing) have the highest and lowest values for light post-editing. The other outputs show a very similar distribution of errors among the two categories of light and full post-editing: In the reference translations a little bit more than half of all errors (~53.9%) require light post-editing. In the output of Tilde, the distribution between the two categories is almost equal. Just a little less than half of the errors (~49.5%) require light post-editing. The output of mBART fine-tuned shows a similar situation: A little bit less than half of all errors (~48.9%) require light post-editing.

To sum up, the reference translations and the output of eTranslation have more errors requiring light post-editing than full. In contrast, the outputs of Tilde, mBART general and mBART fine-tuned have fewer errors requiring light post-editing than full, although one should also keep in mind that, in absolute numbers, there were almost twice as many errors present in the output of mBART general than in the one from eTranslation. While the differences in the number of errors between the two post-editing levels are obvious for the outputs of eTranslation and mBART fine-tuned, they are almost negligible in the reference translations and the outputs of Tilde and mBART fine-tuned.

These results lead to the following two questions asked in the introduction that will consequently help answer the research question:

1) Is the level of light post-editing applied to NMT output from English into Croatian in the medical domain sufficient for content understanding, or is full post-editing required?
2) Does the existing guidance for light vs. full post-editing need to be updated to reflect NMT progress?

Concerning the first question ("Is the level of light post-editing applied to NMT output from English into Croatian in the medical domain sufficient for content understanding, or is full post-editing required?") the answer is clear: According to previously published PE guidelines, light post-editing is not sufficient for content understanding in this domain and for this language pair. All NMT outputs (even the output of the fine-tuned NMT model) contained terminology errors, which is essential for content understanding by the target audience but does not normally come under the light PE requirements. This would suggest that full post-editing is required instead.

However, this leads to the second question: "Does the existing guidance for light vs. full post-editing need to be updated to reflect NMT progress?". Based on the final division of the error categories in this research project into light post-editing and full post-editing in Table 7 (see p. 77) the answer to this question is: Yes, it needs to be updated. As was mentioned before, the guidelines for light post-editing include the following requirements concerning terminology: "No need to spend too much time researching if incorrect" (Mesa-Lao 2013) and "Be consistent" (Densmer 2014). This suggests that incorrect or partially incorrect translations of terminology that require time for researching should not be dealt with in light post-editing. However, terminology errors were among the most frequent error categories in all NMT outputs. The error category *Terminology > Partial error* was the most frequent error category for all outputs except the reference translations. Consequently, terminology errors should be included in light post-editing as well to ensure that the translation is "[f]actually accurate" (Densmer 2014).

As was proposed in this project, *Minor* terminology errors might be ignored in light post-editing as they often do not change the meaning of the source text. However, *Major* terminology errors could indeed change the meaning of the source text, transfer inaccurate information, and thus lead to confusion and misunderstanding. Therefore, *Major* terminology errors should be included in the guidelines for light post-editing. For this proposal to be applicable, the error severity levels should be involved in the updating of the post-editing

guidelines as well. *Critical* errors should always be part of light post-editing alongside *Major* terminology errors.

Considering the updated PE guidelines in this thesis and including *Major* terminology errors, the following proportion of errors in the NMT outputs used in this research project can be corrected with light post-editing: ~38.5% of all errors in the output of mBART general, ~49.5% in the output of Tilde, ~48.9% in the output of mBART fine-tuned and ~62.0% in the output of eTranslation.

Answering these two questions leads to the overall research question of this master's thesis, which is the following: "How applicable is the previously-published guidance for MT light post-editing (produced at the time of statistical MT - SMT) to NMT output of texts from the medical domain for the language pair English-Croatian to enable source-text understanding?" Based on the two previous questions, the following can be concluded: in order to enable source-text understanding, the previously published guidance for MT light post-editing is only applicable to NMT outputs of medical abstracts for the language pair English-Croatian, if it is updated to include *Major* terminology errors as well. These errors endanger the factual accuracy of the text, which is a requirement for target texts going through the process of light post-editing. Therefore, the guidelines for light post-editing are still applicable to NMT output if they are updated to include *Major* terminology errors.

Coming back to the scholarly communication, which is the basis of this research project, the following can be concluded: if Croatian researchers want to use machine translation to translate English abstracts into Croatian to get a summary of research findings, they will need to employ a revised version of existing light post-editing guidance because every NMT output in this project contains errors that need to be addressed to ensure accurate content comprehension in the target language. Light post-editing is especially valuable if short deadlines and a tight budget are the main driving factors. However, if an English abstract should be published in Croatian for a very broad audience in a context where very high quality is needed, full post-editing will still be the better option.

6. Conclusion

The goal of this master's thesis was to analyse the applicability of previously published guidelines for MT post-editing to NMT output in the medical domain. To investigate the chosen research question ("How applicable is the previously-published guidance for MT light post-editing (produced at the time of statistical MT - SMT) to NMT output of texts from the medical domain for the language pair English-Croatian to enable source-text understanding?"), the output of three general NMT models (Tilde, eTranslation and mBART) was analysed alongside a domain-specific fine-tuned version of mBART according to accuracy and fluency against existing reference translations. English abstracts (68 segments from five abstracts in total) from the Croatian scientific magazine *Acta Medica Croatica* served as source texts, and their published Croatian equivalents were used as reference translations.

The first step in the process of this research project included annotating all errors in the NMT outputs with the Qualitivity plug-in for Trados Studio 2021 and assigning them to different error categories. This first step led to the following results:

The comparison between the two NMT models Tilde and eTranslation shows that eTranslation produced more *Accuracy* errors than Tilde, while Tilde produced more *Fluency* errors than eTranslation. eTranslation had more *Major* errors and Tilde more *Minor* errors because *Accuracy* errors are on average rather *Major* and *Fluency* errors are rather *Minor.* Both outputs included two *Critical* errors. In total, both NMT models produced almost the same number of errors.

The comparison between the two NMT models mBART general and mBART fine-tuned shows that mBART general produced both more *Accuracy* and more *Fluency* errors than mBART fine-tuned. mBART general also had both more *Minor* and more *Major* errors than mBART fine-tuned. Moreover, mBART general included more errors in total. Also, by editing the first three categories in the output of mBART fine-tuned, more errors in total can be fixed. Finally, the most frequent *Fluency* error category in the output of mBART fine-tuned is *Mechanical > Typography*, a category that can be edited more easily than others. This means that mBART fine-tuned outperforms mBART general in all aspects.

In contrast to the NMT outputs, the two most frequent *Accuracy* error categories in the reference translations were omissions and additions. Often this does not mean that the reference translations are incorrect, but these types of errors do endanger the status of reference translations as a *gold standard* for annotating errors in NMT outputs. Also, a few terminology errors were found (10 errors) in the reference translations. Considering that these

ten errors are distributed among 68 segments, this does not seem like a high number. Nevertheless, this shows that reference translations produced by human translators do not offer 100% quality. In total, 42 *Fluency* errors were found in the reference translations. Even though all of them are *Minor* errors and therefore "do not impact usability" (Lommel 2018, p. 120-121), they still make clear that the reference translations do not provide 100% quality.

In the second step of this research project, after annotating errors and assigning them to different error categories, these error categories were divided into *light* or *full* post-editing based on a set of post-editing guidelines. These guidelines were compiled from already existing guidelines by TAUS (Massaro et al., 2016), combined with criteria from other guidelines (Hu & Cadwell, 2016). Afterwards it was examined how many error categories and thus errors fall under light and how many under full post-editing. This leads to the following two sub-questions:

1) Is the level of light post-editing applied to NMT output from English into Croatian in the medical domain sufficient for content understanding, or is full post-editing required?

According to previously published PE guidelines, light post-editing is not sufficient for content understanding in this domain and for this language pair. All NMT outputs (even the output of the fine-tuned NMT model) contained terminology errors, which is essential for content understanding by the target audience but does not normally come under the light PE requirements. This would suggest that full post-editing is required instead. However, answering the second sub-question gives a new perspective.

2) Does the existing guidance for light vs. full post-editing need to be updated to reflect NMT progress?

Yes, the existing PE guidance needs to be updated. In the published guidelines for light post-editing the following requirements can be found concerning terminology: "No need to spend too much time researching if incorrect" (Mesa-Lao 2013) and "Be consistent" (Densmer 2014). This suggests that incorrect or partially incorrect translations of terminology that require time for researching should not be dealt with in light post-editing. However, terminology errors were among the most frequent error categories in all NMT outputs. The error category *Terminology > Partial error* was the most frequent error category for all outputs except the reference translations. Consequently, terminology errors should be included in light post-editing as well. As was proposed in this project, *Minor* terminology errors might be ignored in light post-editing as they often do not change the meaning of the source text.

However, *Major* terminology errors could indeed change the meaning of the source text, transfer inaccurate information, and thus lead to confusion and misunderstanding. Therefore, *Major* terminology errors should be included in the guidelines for light post-editing as well.

These two answers lead to the research question examined in this master's thesis: **"How applicable is the previously-published guidance for MT light post-editing (produced at the time of statistical MT - SMT) to NMT output of texts from the medical domain for the language pair English-Croatian to enable source-text understanding?".** The results of this research project show that in order to enable source-text understanding, the previously published guidance for MT light post-editing is only applicable to NMT outputs of medical abstracts for the language pair English-Croatian, if it is updated to include *Major* terminology errors as well. These errors endanger the factual accuracy of the text, which is a requirement for translations going through the process of light post-editing. Therefore, the guidelines for light post-editing are still applicable to NMT output if they are updated to include *Major* terminology errors as well.

Based on the updated PE guidelines in this thesis the following proportion of errors in the NMT outputs used in this research project can be corrected with light post-editing: the output of eTranslation (~62.0% of errors) and mBART general (~38.5% of errors) have the highest and lowest values for errors falling under light post-editing, although in absolute numbers they are in fact comparable (67 errors in the output of eTranslation and 74 errors in the output of mBART general). In the output of Tilde, the distribution between the two categories is almost equal. Just a little less than half of the errors (~49.5% = 53 errors) can be corrected in light post-editing. The output of mBART fine-tuned shows a similar situation: A little bit less than half of all errors (~48.9% = 65 errors) can be corrected in light post-editing. Even the reference translations include errors falling under light post-editing, namely ~53.9% (= 46 errors) of all errors.

Coming back to scholarly communication, which is the basis of this master's thesis, the following can be concluded: if Croatian researchers want to use machine translation to translate English medical abstracts into Croatian to get a summary of research findings, light post-editing in this updated form is feasible. Light post-editing is especially valuable if short deadlines and a tight budget are the main driving factors.

Like all research, this project has a number of limitations. Considering the two main challenges when annotating errors according to Esperança-Rodier (2020), the "selection issue: difference in the number of annotations depending on the typology used" and the "annotation issue: for the same error, the annotator annotated it under one of the typologies and

not the other one" (p. 4), the following limitations need to be mentioned: the author of this thesis was the only annotator in this research project (although, to minimise subjectivity, a gold standard of published translations was used). More annotators are recommended for a similar project in order to discuss into which error category a particular error falls. Also, more segments, as well as more NMT models could be included to gain more data. Moreover, other language directions and language pairs could lead to valuable results.

An approach for further research in this field could be a project focusing on researchers going through MT literacy training to see if they can subsequently modify the source texts to make them more MT-friendly and, in turn, if the target-language researchers can identify any of the remaining MT errors based on source and target MT textual clues.

7. References

Acta medica Croatica, Vol. 75 No. 2, 2021. (n.d.). Retrieved 8 June 2022, from https://hrcak.srce.hr/broj/20695

Acta medica Croatica, Vol. 75 No. 3, 2021. (n.d.). Retrieved 25 May 2022, from https://hrcak.srce.hr/broj/21497

Allen, J. H. (2003). Post-editing. In *Computers and Translation. A translator's guide.: Vol. Volume 35* (pp. 297–317). John Benjamins Publishing Company. https://web-s-ebscohost-com.uaccess.univie.ac.at/ehost/ebookviewer/ebook/bmxlYmtfXzI1MzE5MF9fQU41?sid=6a984fd2-3a79-4c7f-bf22-d881be290821@redis&vid=0&format=EB&rid=1

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Bowker, L., & Ciro, J. B. (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited. https://doi.org/10.1108/9781787567214

Brozović-Rončević, D., Kapetanović, A., & Tadić, M. (2014). *White Paper Serie: The Croatian Language in the Digital Age* (G. Rehm & H. Uszkoreit, Eds.). Springer. https://link.springer.com/book/10.1007/978-3-642-30882-6

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. https://doi.org/10.1515/pralin-2017-0013

Ciobanu, D., Ragni, V., & Secară, A. (2019). Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking. *Informatics*, *6*(4), Article 4. https://doi.org/10.3390/informatics6040051

Daems, J., & Macken, L. (2020). Post-Editing Human Translations and Revising Machine Translations: Impact on efficiency and quality. In *Translation Revision and Post-Editing*. Routledge.

Dunđer, I. (2020). Machine Translation System for the Industry Domain and Croatian Language. *Journal of Information and Organizational Sciences*, *44*, 33–50. https://doi.org/10.31341/jios.44.1.2

Densmer, L. (2014). Light and Full MT Post-Editing Explained. *Moravia*. http://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained

Esperança-Rodier, E. (2020, November). What's on an annotator's mind? Analysis of error typologies to highlight machine translation quality assessment issue. *Translating and the Computer - TC42*. https://hal.archives-ouvertes.fr/hal-03198126

EU Council Presidency Translator. (n.d.). Retrieved 27 June 2022, from https://presidencymt.eu/

Haque, R., Hasanuzzaman, M., & Way, A. (2019). Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 437–446. https://doi.org/10.26615/978-954-452-056-4_052

Hu, K., & Cadwell, P. (2016). A Comparative Study of Post-editing Guidelines. *Baltic Journal of Modern Computing*, *4*(2), 346–353.

ISO. (2015). *ISO 17100:2015: Translation services – Requirements for translation services*. http://www.iso.org/iso/catalogue_detail.htm?csnumber=59149

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. *The Journal of Specialised Translation*, 60–86.

Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2018). Quantitative Fine-Grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, *32*(3), 195–215. https://doi.org/10.1007/s10590-018-9214-x

Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press.

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment* (Vol. 1, pp. 109–127). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_6

Massaro, I., van der Meer, J., O'Brien, S., Hollowood, F., Aranberri, N., & Drescher, K. (2016). *MT Post-Editing Guidelines*. TAUS Signature Editions.

Mesa-Lao, B. (2013). *Introduction to post-editing – The CasMaCat GUI*. http://bridge.cbs.dk/projects/seecat/material/hand-out_post-editing_bmesa-lao.pdf

MQM Council. (2022a). *MQM History—MQM (Multidimensional Quality Metrics)*. https://themqm.org/mqm-history/

MQM Council. (2022b). *Why Use MQM? - MQM (Multidimensional Quality Metrics)*. https://themqm.org/why-use-mqm/

MQM Council. (2022c). *Uses of MQM - MQM (Multidimensional Quality Metrics)*. https://themqm.org/uses-of-mqm/

MQM Council. (2022d). *MQM Core Typology*. https://themqm.info/typology/

Nitzke, J., Hansen-Schirra, S., & Canfora, C. (2018). Risk management and post-editing competence. *JosTrans. The Journal of Specialised Translation*, *31*. https://www.jostrans.org/issue31/art_nitzke.pdf

Nunziatini, M., & Marg, L. (2020). Machine Translation Post-Editing Levels: Breaking Away from the Tradition and Delivering a Tailored Service. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 309–318. https://aclanthology.org/2020.eamt-1.33

O'Brien, S. (2010, October 31). Introduction to Post-Editing: Who, What, How and Where to Next? *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Tutorials*. https://aclanthology.org/2010.amta-tutorials.1

Presidency MT. (n.d.). Retrieved 14 October 2022, from http://www.tilde.com/products-and-services/machine-translation/de-presidency

Prokopidis, P., & Papavassiliou, V. (2020, February). *Multilingual corpus made out of PDF documents from the European Medicines Agency (EMEA)*. European Language Resource Coordination. https://elrc-share.eu/repository/browse/multilingual-corpus-made-out-of-pdf-documents-from-the-european-medicines-agency-emea-httpswwwemaeuropaeu-february-2020/3cf9da8e858511ea913100155d0267062d01c2d847c349628584d10293948de3/

Qualitivity. (n.d.). RWS AppStore > Wiki. Retrieved 9 June 2022, from https://community.rws.com/product-groups/trados-portfolio/rws-appstore/w/wiki/2251/qualitivity

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., & Fan, A. (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning* (arXiv:2008.00401). arXiv. http://arxiv.org/abs/2008.00401

Van Brussel, L., Tezcan, A., & Macken, L. (2018, May). A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1600

von Platen, P. (n.d.). *MBart and MBart-50*. Retrieved 25 June 2022, from https://huggingface.co/docs/transformers/model_doc/mbart

Way, A. (2018). Quality Expectations of Machine Translation. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 159–178). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7

8. List of figures

9. List of tables

10. Appendix

Definitions of the general categories for error annotation by MQM (MQM Council, 2022a):

Accuracy

Accuracy > Mistranslation

The target content does not accurately represent the source content.

Example: A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).

Accuracy > Omission

Content is missing from the translation that is present in the source.

Example: A paragraph present in the source is missing in the translation.

Accuracy > Addition

Target content that includes content not present in the source.

Example: A translation includes portions of another translation that were inadvertently pasted into the document.

Accuracy > Terminology

A term (domain-specific word) is translated with a term other than the one expected for the domain or otherwise specified.

Example: A French text translates English e-mail as e-mail but terminology guidelines mandated that courriel be used.

Accuracy > Untranslated

Content that should have been translated has been left untranslated.

Example: A sentence in a Japanese document translated into English is left in Japanese.

Fluency

Fluency > Content > Inconsistency

The text shows internal inconsistency.

Example: The text states that bug reports should be submitted to a mailing list in one place and via an online bug tracker tool in another.

Fluency > Content > Register

The text uses a linguistic register inconsistent with the specifications or general language conventions.

Example: A legal notice in German uses the informal du instead of the formal Sie.

Fluency > Content > Stylistics

The text has stylistic problems, other than those related to language register.

Example: A text uses a confusing style with long sentences that are difficult to understand.

Fluency > Mechanical > Grammar

Issues related to the grammar or syntax of the text, other than spelling and orthography.

Example: An English text reads "The man was seeing the his wife."

Fluency > Mechanical > Locale convention

The text does not adhere to locale-specific mechanical conventions and violates requirements for the presentation of content in the target locale.

Example: An incorrect format for currency is used for a German text, with a period (.) instead of a comma (,) as a thousands separator.

Notes: This issue type is distinguished from locale-specific-content in that this category refers only to whether the text is given the proper mechanical form for the locale, not whether the content applies to the locale or not. If text conforms to conventions for the locale, but does not apply to the target locale, locale-specific-content should be used instead.

Fluency > Mechanical > Spelling

Issues related to spelling of words.

Example: The German word Zustellung is spelled Zustetlugn.

Fluency > Mechanical > Style guide

The text violates style defined in a normative style specification.

Example: Specifications stated that English text was to be formatted according to the Chicago Manual of Style, but the text delivered followed the American Psychological Association style guide.

Fluency > Mechanical > Typography

Issues related to the mechanical presentation of text. This category should be used for any typographical errors other than spelling.

Example: A text uses punctuation incorrectly.

Notes: Do not use for issues related to spelling.

Fluency > Unintelligible

The exact nature of the error cannot be determined. Indicates a major break down in fluency.

Example: The following text appears in an English translation of a German automotive manual: "The brake from whe this કુતારો િસ S149235 part numbr,,."

Verity > Completeness

The text is incomplete.

Example: A process description leaves out key steps needed to complete the process, resulting in an incomplete description of the process.

Verity > Legal requirements

A text does not meet legal requirements as set forth in the specifications.

Example: Specifications stated that FCC regulatory notices be replaced by CE notices rather than translated, but they were translated instead, rendering the text legally problematic for use in Europe.

Verity > Locale-specific content

Content specific to the source locale does not apply to the intended target locale, audience, or purpose.

Example: An advertising text translated for Sweden refers to special offers available only in Germany and therefore is misleading.

Notes: This issue type is distinguished from locale-convention in that this category applies to cases where text corresponds to the conventions of the target locale but does not apply to the intended audience in the target locale. For example, if the Swedish advertising text mentioned above is properly translated and follows all mechanical locale conventions (e.g., using Swedish kronor instead of euros) but the offer does not apply to Sweden, locale-specific-content should be chosen. If, however, the text applies to the locale, but does not follow locale conventions (e.g., numbers are formatted incorrectly for the locale), locale-convention should be used instead.

Definitions of the terminology categories for error annotation (Haque et al., 2019):

Accuracy > Terminology > xxx

Reorder error:

the translation of a source term forms the wrong word order in target;

Inflectional error:

the translation of a source term inflicts a morphological

error (e.g. includes an inflectional morpheme that misfits in the context of the

target translation, which essentially causes a grammatical error in translation);

Partial error:

the MT system correctly translates part of a source term into target and commits an error for the remainder of the source term;

Incorrect lexical selection:

the translation of a source term is an incorrect

lexical choice;

Term drop:

the MT system omits the source term in translation;

Source term copied:

a source term or part of it is copied verbatim to target;

Disambiguation issue in target:

although the MT system makes a potentially correct lexical choice for a source term, its translation-equivalent does not carry the meaning of the source term. In other words, its sense is different to the source term;

Other error:

there is an error in relation to the translation of a source.

Abstract

English

The medical domain is characterised by intense international cooperation, leading to scientific publications being authored mostly in English. With short deadlines and a tight budget, a professional translation of scientific papers or abstracts is often not an option. In this case, machine translation combined with post-editing can be a very useful tool to make research findings available quicker to a broader audience. This master's thesis analyses how applicable the previously published guidance for MT light post-editing (produced at the time of statistical MT) is to ensure the accurate understanding of NMT output of texts from the medical domain for the language pair English-Croatian. 68 segments taken from English medical abstracts are machine-translated by three general NMT models (Tilde, eTranslation and mBART) alongside a domain-specific fine-tuned version of mBART. The errors in the MT outputs are annotated using the Qualitivity plug-in from Trados Studio 2021 and compared to guidelines for light and full post-editing (PE) by the language data network TAUS (Massaro et al., 2016), combined with criteria from other guidelines (Hu & Cadwell, 2016). The results of this research project show that the previously published guidance for MT light post-editing is still applicable to NMT outputs of medical abstracts for the language pair English-Croatian in an updated format. In order to meet the comprehensibility needs of the intended target audience of these abstracts, the existing PE guidelines were updated to include terminology errors with the error severity *Major* in light post-editing, as well. As a result, the following proportion of errors in the NMT outputs used in this research project can be corrected with light post-editing: Tilde ~49.5%, eTranslation ~62.0%, mBART general ~38.5% of and mBART fine-tuned ~48.9% of all errors.

Deutsch

Der medizinische Fachbereich ist durch intensive internationale Zusammenarbeit gekennzeichnet, bei der wissenschaftliche Publikationen häufig in englischer Sprache verfasst werden. Knappe Deadlines und ein begrenztes Budget machen eine professionelle Übersetzung wissenschaftlicher Arbeiten oder Abstracts oft unmöglich. In diesen Fällen kann maschinelles Übersetzen zusammen mit Post-Editing eine mögliche Lösung sein, um Forschungsergebnisse einem breiteren Publikum schneller zugänglich zu machen. In dieser Masterarbeit wird untersucht, inwieweit die bisher veröffentlichten Richtlinien für light Post-Editing (die zur Zeit der statistischen maschinellen Übersetzung erstellt wurden) richtiges Verständnis neuronaler maschineller Übersetzungen (NMT) von Texten aus dem medizinischen Bereich für das Sprachenpaar Englisch-Kroatisch sicherstellen. 68 Segmente aus medizinischen Abstracts (verfasst auf Englisch) wurden mit drei allgemeinen NMT-Modellen (Tilde, eTranslation und mBART) sowie einer domänenspezifischen, auf medizinische Texten trainierten Version von mBART maschinell übersetzt. Die Fehler in den maschinell übersetzten Texten wurden mit dem Qualitivity-Plug-in von Trados Studio 2021 annotiert und mit den Richtlinien für light und full Post-Editing (PE) des language data networks TAUS (Massaro et al., 2016), kombiniert mit Kriterien aus anderen Richtlinien (Hu & Cadwell, 2016), verglichen. Die Ergebnisse dieser Masterarbeit zeigen, dass die zuvor veröffentlichten Richtlinien für light Post-Editing in aktualisierter Form auf neuronale maschinelle Übersetzungen von medizinischen Abstracts für das Sprachpaar Englisch-Kroatisch anwendbar sind. Um den Verständlichkeitsanforderungen der Zielgruppe der medizinischen Abstracts gerecht zu werden, wurden die bestehenden PE-Richtlinien aktualisiert, um auch Terminologiefehler mit dem Fehlerschweregrad *Major* in light Post-Editing zu inkludieren. Infolgedessen kann der folgende Anteil an Fehlern in den neuronal maschinell erstellten Übersetzungen dieses Projektes durch light Post-Editing korrigiert werden: Tilde ~49,5%, eTranslation ~62,0%, mBART general ~38,5% und mBART fine-tuned ~48,9% aller Fehler.