The Use of Neural Machine Translation in Translating Finnish News Articles – An Error Analysis of the NMT Service DeepL

Samu Lindvall
Bachelor's Thesis
682285A Bachelor's Seminar and Thesis
English
Language and Literature
Faculty of Humanities
University of Oulu
Spring 2022

Abstract

In this thesis, a brief overview of the functionality of the neural machine translation system DeepL is provided. Machine translation is an expanding field in translation studies, and it continuously provides us with new technology and applications to translate texts as accurately as possible. The aim of this study was to examine the capabilities of DeepL in translating Finnish news articles with no available reference translations. The reason for this was to prevent DeepL from finding completed translations and possibly benefitting from them, as neural machine translation systems seek data from sources available on the internet. The articles were selected randomly from the free internet news providers Yle and Iltalehti. The errors have been listed, categorized and analyzed in the section "Analysis". Conclusions along with general discussion about the performance of DeepL can be found in the last section of this thesis. Overall, this thesis shows DeepL's promising capability when translating news text. Still, it must be kept in mind that the aim of this thesis was not to seek perfect translations, but rather successful message transmission. The idea of machine translation being a worthy competitor to human-made translations in more specific areas of translation, such as medical or legal translation, is still far away. Still, any conducted research is vital for the progression and development of machine translation services. The analysis of this study provides examples of areas where DeepL is not sufficient. These areas include for example, the translation of new words, translation of pronouns and culture-specific terms. Instances when DeepL succeeds to make acceptable translations in one of the listed categories, have been presented also in the analysis section.

Tiivistelmä

Kandidaatintyössäni käsitellään neuroverkkoihin perustuvan käännösohjelman, DeepL:n suoriutumista käännettäessä suomalaisia uutistekstejä. Käytettävät tekstit on poimittu Ylen ja Iltalehden ilmaisista nettiuutispalveluista. Työssä käytettyjä uutistekstejä ei ole käännetty englannin kielelle. Tämä oletettavasti välttää konekääntöohjelmien hyötymisen valmiista käännöksistä, joka on tärkeää ottaen huomioon neuroverkkoihin perustuvien kääntämisohjelmien toimintaperusteet.

Konekääntäminen on aiheena trendikäs, ja kuuluu nykypäivänä sovellusten myötä osaksi lähes jokaisen elämää. Alan kehittymisen voi selvästi havaita tarkkailemalla koko ajan uudistuvia konekääntöohjelmia. Käännetyistä teksteistä havaitut virheet on analysoitu virheanalyysin muodossa, ja johtopäätökset esitetty tutkielman lopussa. Virheanalyysi on jaettu kategorioihin virhetyyppien perusteella. Yleisellä tasolla voinee todeta, että DeepL suoriutuu hyvin käännettäessä suomalaista uutistekstiä. Tämän tutkimuksen tavoitteena ei kuitenkaan ollut etsiä täydellistä kääntämistä, vaan onnistunutta viestinvälitystä. Eri kääntämisen alat, kuten lääketieteellinen kääntäminen ja lakitekstikääntäminen vaativat äärimmäistä tarkkuutta, ja tässä työssä esitettävien virhe-esimerkkien perusteella voinee todeta, että konekääntäminen ei vielä sovellu vaikkapa edellä mainittujen alojen tekstien kääntämiseen. Konekääntäminen on joka tapauksessa toimiva apuväline jokapäiväisiin kieleen liittyviin ongelmiin, ja sen tutkiminen on tärkeää sen kehittämiselle. Analyysini antaa pintaraapaisun siitä, missä DeepL:n kaltaiset konekääntöohjelmat eivät vielä suoriudu. Virheitä löytyi esimerkiksi pronominien käytössä, kulttuurille omien termien käännöksessä ja uudissanojen kääntämisessä. Analyysissä on esitelty myös esimerkkejä tapauksista, joissa DeepL suoriutuu kääntämään tiettyyn kategoriaan liittyvän tekstin osan onnistuneesti.

Table of Contents

1. Introduction

In today's interconnected world we often find ourselves in a situation where information we are trying to gather is not available in our own language. Luckily, in the era of computers and smart devices, translation services are within reach of us all. Machine translation (henceforth MT) is a constantly growing and improving field in translation studies, as well as the core of the many online translation services we come across today. Mostly thanks to the global push for automatization, its importance on the field has greatly increased. The almost endless sources of new information, social media and news outlets to name a few, provide us with new content in such speed and quantity that it makes the traditional human methods of translation an overwhelming task. However, it is important to keep in mind that this only applies to a specific group of texts and that, there is still a time and place for human translation, or as Way (2018, p.2) points out: "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."

The subject of this study, neural machine translation (henceforth NMT) is a relatively new subbranch of MT, with first studies introducing the approach being published in 2013-2014 (Kalchbrenner & Blunsom, 2013; Sutskever et al., 2014). NMT revolutionized the field of MT with its accuracy (Kalchbrenner & Blunsom, 2013) and lower demand for data (e.g. Cho et al., 2014b). The most widely known NMT service, Google Translate, along with many others, are used for a wide variety of tasks by millions of people every day (Way, 2018).

In this study I will be examining the NMT service DeepL. Similarly to its competitors in the field, it too uses neural networks to translate text. However, DeepL claims to provide users with better results due to added improvements on network architecture, training data, training methodology and network size (DeepL, 2021). With examining accuracy, terminology, fluency and verity with manually categorizing errors, and providing examples of occasions where NMT positively stands out, I will present a brief outlook of the overall proficiency of DeepL in translating news articles in the Finnish language. As a resource, news articles provide challenges to NMT, but simultaneously, it is among those translation categories that I would see to benefit the most from NMT. In fact, according to Hassan et al (2018, p.1), "translation of news text has been an area of active interest in the

Machine Translation community for over a decade, due to the practical and commercial importance of this domain."

The field of machine translation is vast and includes different approaches to translate texts from one language to another. Since these methods used to construct machine translation systems are technical and usually include implementations of from the fields of mathematics and computational linguistics, I find that it is necessary to cover the terminology and define both the basic features and methodology of the translation services mentioned in this study.

Machine translation (MT) is an umbrella term for the expanding field of computer software translation systems. Hutchins (2007, p.1) summarizes three usages for MT in general as follows: Firstly, MT can be used to produce "draft translations" which are then later edited and published. Secondly, MT can be used to "convey the essence" of the original text. In this context, essence is used to symbolize the core meaning of the original text. Thirdly, in more recent applications, MT can be used to aid communication between speakers of different languages. For example, Google Translate is able to translate speech in real time through their mobile application. The latter two usages do not aim for perfect, publishable text, whereas the first-mentioned usage (with help of either automated or human-carried methods) aims for high quality publishable texts.

Although the development of MT began as early as 1933 (Hutchins, 2007), this study will be focusing on the more recent MT approaches, in this case NMT. To illustrate how far the technology has improved and to better understand NMT itself, a brief historical outline is provided. Without getting into details, the historical development of MT can be roughly categorized into three stages chronologically: firstly, to the era of rule-based machine translation, where syntactic, lexical, and morphological rules were programmed to an application and used to translate texts, followed by the era of corpus-based machine translation, where mathematic, statistical methods and parallel corpora were used to translate texts, and more recently, neural machine translation. (Hutchins, 2007)

The currently dominating paradigm (Rivera-Trigueros, 2021), neural machine translation, NMT, differs from past paradigms in the way that its aim is to build a "single, large neural network that reads a sentence and outputs a correct translation." (Bahdanau et al., 2015, p.1) NMT translation services like DeepL typically use the *encoder-decoder-method*, which means that the source text is first encoded to a fixed-length decodable vector which is then decoded with a language-specific

decoder. These encoders and decoders are trained for each language pair specifically (Bahdanau et al., 2015). In simpler terms, a source sentence in L1 is encoded to a fixed-length decodable "code" (vector) which can then be decoded with a model made from L2, resulting in a translated sentence of varying length. For both the corpus-based, still widely used statistical machine translation (SMT) and NMT, the data is gathered with automatic extraction of linguistic corpora. These corpora can be found from websites, e-books and other sources. (Rikters, 2018). There however a major difference in the operating principle between these paradigms. SMT statistically analyzes corpora in real time to come up with the probability of a translation, while for NMT, the *encoder-decoder-system* only has to be trained once with the data, resulting in NMT demanding a fraction of the memory that SMT does (Cho et al., 2014b).

2. Background and analytical framework

Research on NMT ranges from purely mathematical studies, which present new approaches and evaluate the capability of older ones, to research done in the field of humanities, in which authors explore questions related to language and linguistics. Different translation services use these new approaches to design new, better systems. It is also a common theme in MT research to compare the capability of two or more translation services. Cambedda et al. (2021) conducted a study using comparative analysis, comparing the results of NMT services DeepL and Yandex with special attention to medical translations. They both manually detect errors via linguistic criteria as well as evaluate the translation programs with the BLEU metric, which is an automated evaluation method that gives machine-generated translations a score between 0 and 1 by comparing the occurrence rates of matching words between candidate and reference translations (Papineni, 2002). They found that DeepL provided slightly higher quality translations than its competitor, but that the translations in general were not accurate and natural enough and that the field demands more work (Cambedda et al., 2021). Evaluation of the produced translations seems to also be a hot topic in the field of MT in general (see e.g. Rivera-Trigueros, 2021; Farrús et al., 2010; Graham et al., 2015; Kim et al., 2020). New and existing methodology is also evaluated and new ideologies are presented in the aim for an efficient approach to evaluating machine translations. In general, many studies focus on the problematics of translation quality assessment, as well as on the fact that there is no clear consensus in the field of MT assessment (Rivera-Trigueros, 2021).

The manual analysis of errors, which is the methodology used in this study, has both benefits and disadvantages. As presented by Rikters (2018), both MT systems, statistical and neural, rely on vast parallel data corpora, which will most probably cause errors. However, with manual assessment, these errors caused by parallel data can perhaps be recognized better than with automated processes. As noted by Rivera-Trigueros (2021, p.6), "automatic methods – give rise to confusion between the different types of errors – and they also require human reference translations." This applies to the most used automated evaluation systems, like BLEU, defined above. The materials of this study do not yet have any reference translations, which could make the usage of automated evaluation systems like BLEU challenging. Although BLEU has been proven to be correlate with human judgement (Papineni, 2002), new, emerging texts that need to be translated quickly do not have reference translations in order to compare the resulted translation with a finished translation. Another benefit for the manual assessment of errors with linguistic guidelines could be that errors can be categorized clearly. For example, Farrús et al. (2010) present the following five (5) categories of errors; orthographic, morphological, lexical, semantic and syntactic. Naturally, this is only one example of how MT errors can be categorized or assessed. Although recommendations of the usage of both automated and human systems for assessing errors exist (Chatzikoumi, 2020; Way, 2018), automated systems like BLEU still dominate the error assessment field (Castilho et al., 2018). Disadvantages for the type of approach that will be used for this study naturally include that it is time-consuming and usually presents problems with the inter-annotator agreement (Rivera-Trigueros, 2021), which will be defined later in this chapter.

The news articles used for this study are translated without picture captions with the translator found on DeepL's website. Although DeepL suggests alternative translations for each word, only the first suggested translation is taken into account. In other words, we assume that the user is not familiar with L1. In order to analyze the resulting translations, a custom manual assessment category set is created with the guidelines provided by Deutsches Forschungszentrum für Künstliche Intelligenz or German Research Center for Artificial Intelligence (DFKI) (2015). Their Multidimensional Quality Metrics (MQM) is a free-to-use framework for translation quality assessment. With MQM, the user can create a custom translation evaluation metric, with the help of the provided error types (QT21, 2022). It should be noted, that as stated in the copyright notation for MQM, the categorization system used in this thesis is created by the author of this thesis, and that it does not represent DFKI or have its endorsement in any shape or form. In MQM, translation

quality is divided into seven main categories (accuracy, fluency, design, locale convention, style, terminology and verity) with a varying amount of subcategories for each group. With this metric, the errors noticed in the data can be categorized accordingly. The categorization method was formulated based on the errors noticed when manually assessing the texts, and it does not include all the categories present in MQM. With the added definition from DFKI (2015) for each error category, the analysis section can be observed by a reader not familiar with MQM. In addition to categorizing, hypotheses for the cause of the errors are provided for some examples, based on the basic functions of NMT. Since my aim is to evaluate the overall performance of DeepL in translating news articles, some examples of correct translations are also provided. An assumption about the proficiency of the reader is made for this study. We assume that the reader obtains a proficiency of English to the point that clear errors regarding intelligibility can be easily noted by the reader. Still, for this thesis, a clear emphasis is placed on "conveying the essence" (Hutchins, 2007).

It should also be noted that the inter-annotator agreement, or the "measure of how well two (or more) annotators can make the same annotation decision for a certain category," (Corpus Linguistic Methods, 2014) has to be kept in mind while examining the results. Although one could argue that there is no single correct translation, the results of this study are highly subjective. Still, the results provide a brief introduction to what we can expect from NMT services like DeepL.

3. Materials

The main materials for this study are news articles written in the Finnish language. The texts are specifically picked so that they have no existing reference translations. This is because although the neural networks used by my target translation service, DeepL, are trained excessively with various text corpora to produce human-like translations, they could still benefit from existing translations of a news article in question because of the use of highly intelligent *web-crawlers* (DeepL, 2021), a form of automatic corpora extraction. News articles are particularly tricky in a sense that they are incredibly diverse; all authors have their own style of writing and might sprinkle idioms and nuances to the articles to make them more appealing to the reader.

The articles are from free sources only. No text used in this study requires a paid subscription for usage, mainly because of ethical and copyright issues. In general, the nature of this study does not

present any ethical issues due to the materials being public and free of use. The MQM presented by DFKI (2015) is also a free-to-use service, and the copyright clauses have been assessed in the previous chapter of this study. The used articles differ in category, including, for example, sports news and articles about the COVID-19 pandemic.

The main service used for this study is Yle, a national government-owned media service, that provides news services in both TV and the internet. Yle.fi has its own section, "News", which provides its English-speaking readers with professionally translated articles, but naturally, the selection is very limited compared to the Finnish articles. However, the articles mentioned in this thesis did not have any reference translations in Yle News at the time of the analysis. It seemed to be a reoccurring phenomenon that only the most important news articles were translated into English. Iltalehti is the other news service used in this study. It is a Finnish tabloid newspaper company owned by Alma Media Suomi Oy. Iltalehti has no specific section written in the English language, which makes it an excellent candidate for this study.

In total, 12 articles were analyzed, nine from Yle and three from Iltalehti. Clear issues in translations were found from nine of these articles. The source of the article seemed not to reflect the possibility of errors, as errors were found in articles from both sources.

4. Analysis

The analysis section of this study is split into the following parts: Accuracy, Terminology, Fluency and Verity. These parts are formulated followingly: After analyzing the translations and manually picking out errors, these errors are then categorized according to the guidelines presented by DFKI (2015). As mentioned before, this categorization is created by the author of this thesis, and these examples are not approved nor endorsed by DFKI. At the start of every category, an official definition is provided. The examples include further discussion about the nature of how and why this mistake may have occurred. It is also to be noted that the term *source text* always refers to the original Finnish article, and *target text* refers to the resulted translation. All the translated texts can be found in a list placed under the references for this study. Some of the texts were used solely to gather more data, and therefore the analysis does not include examples of them. Also, some of the analyzed texts did not present any clear issues.

4.1 Accuracy

Accuracy (or adequacy) of a translation evaluates whether the target text reflects the source text (DFKI, 2015). In other words, this chapter evaluates whether the message of the source text is sufficiently translated to the target text. In this category, mistranslations, additions, overly literal translations and omissions are evaluated. DFKI, 2015, explains additions and omissions as follows: Addition means that the resulted translation includes text not present in the original text. Omission means that the translation does not include a vital part of the source text. The type of error is presented in the first line of each example. Most of the noticed errors belonged to this category.

(1) Error: Addition

Original article by Muilu, Hannele (16.1.2022, Yle). Article topic: The major problem of workers being on sick leave or quarantined. Covid-19-related. This same article is used also in Examples 3 and 5.

pidämme täällä hyvin maskeja ja turvavälejä [original Finnish text]
we keep masks and safety goggles well here [DeepL initial translation]

In this example, we can clearly notice that the resulting translation includes words that were not present in the original text. The word *Turvalasit* [safety goggles] does not exist in the source text. This mistake could have occurred due to word-prediction. *Turvaväli* [social distancing/distance] is translated to *safety goggles,* most probably because of the neighboring word *masks.* DeepL does not understand the type of masks used, and most likely refers to respirators used in construction zones, hence the word *safety goggles*.

(2) Error: Mistranslation – Overly literal

Original article by Ketonen, Petra (12.5.2022). Article topic: The over-usage of occupational health care services and its effects to work disability. This article is also used in examples 4, 6 and 7.

työskentelivät usein suorittavassa työssä [original Finnish text]
often worked in performing jobs [DeepL initial translation]

In this example, we can see an overly literal translation of the Finnish word *suorittava työ* [manual labor]. As DeepL is not able to correctly recognize the term, it translates the word *suorittava* into *performing,* resulting in a clear ambiguity, because the reader might think that the article is discussing occupations that have to do with performing or arts and crafts.

(3) Error: Mistranslation – Overly literal

Original article by Muilu, Hannele (16.1.2022, Yle).

Luulen, että kuminauhaa ei sieltä löydy [original Finnish text]
I don't think there will be a rubber band [DeepL initial translation]

Symbolism is a tough linguistical characteristic for DeepL. In this sentence, *kuminauha* [rubber band] is used to symbolize flexibility. The correct translation for this sentence would be "I don't think there will be any flexibility." Symbols differ from idioms in a sense that there are no fixed expressions for them, as almost any word can be used to symbolize the quality of another. This results in an overly literal translation of the word *kuminauha.*

(4) Error: Addition

Original article by Ketonen, Petra (12.5.2022).

mutta haitari on todella iso [original Finnish text]
but it's a really big number [DeepL initial translation]

In Example 4, DeepL is able to successfully recognize the context of the text, but still does not succeed in providing a correct translation. In the original article, the Finnish word *haitari* [accordion] was used to symbolize the variability of a number, ranging from 0 to 8. The resulted translation is a clear error, but on the other hand shows that DeepL is capable of recognizing contexts where a single word should not be translated literally.

(5) Successful translation from one idiom in L2 to another one in L2

Original article by Muilu, Hannele (16.1.2022, Yle).

vaikeampi pala nieltäväksi [original Finnish text]
harder pill to swallow [DeepL initial translation]

In this Example 5, we can see that DeepL is capable of recognizing more widely known idioms, resulting in a perfect translation from one idiom to another. Although the words *pala* [piece] and *pill* do not have anything in common, the trained networks replace the words with each other to create a functioning idiom in L2.

(6) Error: Omission

Original article by Ketonen, Petra (12.5.2022).

puuttuminen on kuitenkin hyvin vaikeaa [original Finnish text]
it is very difficult to avoid [DeepL initial translation]

In Example 6, DeepL tries to paraphrase the meaning of the original sentence. Although the word *puuttuminen* [intervening] is omitted from the translation, the core meaning of the utterance stays almost the same. However, as the translation does not include the word *intervene,* it should be classified as an error, since it is a vital part of the original sentence.

(7) Error: Overly literal / addition

Original article by Ketonen, Petra (12.5.2022).

terveysjäteille [original Finnish text]
into big health garbage [DeepL initial translation]

In Example 7, we can see an overly literal translation of the word *terveysjätti* [healthcare giant]. The resulting translation, *big health garbage,* has most likely resulted from a falsely recognized typo, since the words inflected forms of the Finnish words *jätti* [giant] and *jäte* [rubbish, garbage] are similar to each other. In the analysis section, discussion about the possibility of DeepL falsely correcting presumable typos can also be found in Example 8. An addition of a single letter *t* would change the original word to *terveysjätteille*, which would make the translation nearly correct. Still, the added word *big* along with a clear mistranslation makes the example unintelligible.

Overall in this first error category, we can see that mistranslations belonging to the listed subcategories usually occur when DeepL has to deal with words that are used to symbolize other words, or words used as a part of an locally used idiom. DeepL also struggles with new words and terminology connected to them. This phenomenon will be discussed further in the next error category. In the examples in this first category, word prediction, a core feature of NMT translations services, can actually prove to be disadvantageous to the translation.

4.2 Terminology

The following examples presented belong to the category *Terminology.* In this category, the accuracy of translated terminology is assessed (DFKI, 2015). The type of error is presented in the first line of each example.

(8) Error: Inconsistent use of terminology – Multiple translations of same term / (Overly literal)

Original article by Pikkarainen, Aleksanteri (25.1.2022, Iltalehti). Article topic: Covid-related restrictions and the statements of the Finland's current Prime minister Sanna Marin. This text is used also in Example 10.

koronapassi [original Finnish text]
interest rate pass, (coupon pass) [DeepL initial translation]

In Example 8, DeepL is not able to recognize the word *koronapassi* [COVID vaccination pass] but rather translates the word to *interest rate pass,* which is due to the Finnish word *korko* meaning interest rate. Interestingly enough, DeepL *is able* to translate the word by itself, but when used in a sentence, the word changes with regard to the other words used in the same sentence, or the sentences surrounding it. The original Finnish sentence "Pääministeri Sanna Marin (sd) kertoo

Twitterissä, että matalariskisten kulttuuri- ja urheilutilaisuuksien avaamista koronapassilla sekä ruokaravintoloiden aukiolon laajentamista pitää harkita arvioitua nopeammin" [Prime minister Sanna Marin informs on Twitter that both the allowance for low-risk cultural and sporting events and expanding opening hours of restaurants needs to be considered earlier than expected] has 2 utterances that affect the resulting translation*, low-risk* and *cultural and sporting events*. If both are removed, the translation is correct. *Low-risk* changes the word to *interest rate pass. Cultural and sporting events* changes the word to *coupon pass*. It should be noted that these two words are clear mistranslations, which means this example arguably could belong to the category "Accuracy", but as DeepL is not consistently able to translate the term itself, this example is discussed under this section. However, the above is a concrete example of how MT predicts words based on context.

When considering this example further, we can come up with another possible reason for this mistranslation. In an earlier example, we talked about the possibility of DeepL falsely recognizing typos and correcting them in order to make a successful translation. For this example, that could also be the case. If the letter *a* is omitted from the original word *koronapassi* and considered as a typo, the word *koronpassi* could be translated to *interest rate's pass*. Both of these words are nonetheless both unintelligible and unidiomatic.

(9) Error: Inconsistent use of terminology – Multiple translations of same term / (Addition)

Original article by Taleva, Katariina (16.5.2022, Iltalehti). Article topic: The abandonment of pets acquired during the time of the pandemic. This article is also used in Example 15.

korona-aika [original Finnish text]
Crown era [DeepL initial translation – 1st sentence]
loan period [2nd sentence]
the end of the basking season [3rd sentence]

In this example we can witness extreme inconsistency on translating terms related to the COVID-19 pandemic. *Korona-aika* [COVID-era] is translated into three different terms, *crown era, loan period* and *the end of the basking season,* which are all completely incorrect and cause major faults regarding the accuracy of the translation. Interestingly enough, these translations happen one after another, namely in the first three sentences of the original article. These sentences include the heading, subheading and the first sentence of the article. This example is also closely connected to Example 8, due to the similar words presented by the translator. The original sentences of the article were similar and contained mostly similar words, only in a different order. The removal of neighboring words seemed to not affect the resulting translation. This example shows how DeepL is not yet familiar with recently emerged new words, and creates translations that would most likely cause confusion.

The overall findings of this chapter can be summarized to the following: DeepL is not able to adequately translate newly emerged words, in this case terminology connected to the COVID-19 pandemic. These words are new and should be safe to say that there is still probably an insufficient amount of multilingual corpora for DeepL to be able to sufficiently translate these terms, their variants and inflected forms.

4.3 Fluency

Fluency deals with the form of the resulted translations. Grammatical aspects, such as pronoun usage and word form are considered within this chapter (DFKI, 2015). Although the aim of this study is not to focus solely on formality or absolute correctness of a text, some errors give rise to general confusion, and therefore have to be addressed. When translating a group of texts, it was evident that for DeepL has some clear issues in recognizing gender from context. Therefore, issues related to gender will be discussed in this chapter. However, for the focus point of this study, DeepL succeeded in providing grammatically accurate translations, which will be discussed in the section *Conclusion*.

DeepL seems to generally handle the Finnish gender-neutral pronoun *hän* [he/she]. Out of a smaller subset of 26 occasions of the pronoun in inflected or non-inflected forms in target texts, including two occasions where the gender was unknown, only 5 clear mistranslations occurred. In these cases,

DeepL tended to favor the pronoun *he*, switching into it immediately if the person's name was not in the pronoun's immediate vicinity, another proof on how NMT translators carefully examine words presumably only in the closest sentences. Unfortunately, examining only the neighboring sentences in a text is not enough, as mistranslations tend to happen when a name presented earlier appears again in the text. DeepL was very sufficient in recognizing gender from Finnish first names if the name was included in the sentence. Examples of this phenomenon can be found in the following chapter. The issue type can be found in the first line of each example. Successful translations are also provided in during the chapter.

(10) Error: Grammar

Original article by Pikkarainen, Aleksanteri (25.1.2022, Iltalehti).

Pääministeri Sanna Marin kertoo . . . hän jatkaa. [original Finnish text]
Prime Minister Sanna Marin says . . . he continues. [DeepL initial translation]

DeepL is not able to recognize that the operator of the text has remained the same, and suggests the pronoun *he* to be used when addressing Sanna Marin, a female Prime Minister.

(11) Error: Grammar

Original article by Eskonen, Hanna (18.5.2022) Article topic: The possibility of rising costs in the public health care system due to inequalities in pay. This article is also used in Example 16.

Satu Ojala [original Finnish text]
Mr. Ojala [DeepL initial translation]
he [2nd sentence]
she [3rd sentence]

DeepL showed clear inconsistency in referring to a common Finnish name, *Satu*. The translations followed the pattern described earlier, a correct translation occurring when the name was in a neighboring sentence, and an incorrect translation (*he, Mr.*) occurring when the name was unknown in the context.

(12) Successfully translated pronoun

Original article by Hallamaa, Teemu (6.2.2022, Yle). Article topic: Facebook and content sharing in Finland. This article is also used in Example 13.

Käyttäjällä ei ollut mahdollisuutta ymmärtää, minkälaisen ketjun päätteeksi julkaisu oli päätynyt hänen ruudulleen. [original Finnish text]
Users had no way of knowing which chain had led to the publication reaching their screen [DeepL initial translation]

In Example 12, DeepL provides a correct translation of the gender neutrally used pronoun *hän,* translating it to *their.*

(13) Successfully translated pronoun

Original article by Hallamaa, Teemu (6.2.2022, Yle).

esimerkiksi kysymällä käyttäjältä, onko tämä lukenut kyseessä olevan artikkelin [original Finnish text]
for example by asking a user if he or she has read the article in question [DeepL initial translation]

In Example 13, the Finnish pronoun *tämä* is used to refer to the last operator mentioned in the text, in this case, the user. DeepL successfully translates it neutrally, resulting in the translation *he or she.*

To conclude this section of the analysis, the translation of Finnish names and pronouns seems to not be a major problem to DeepL. However, this analysis having been done using only Finnish texts and only using common names, we can not draw any major conclusions on how DeepL handles names and pronouns. It is clear however that DeepL does not sufficiently translate pronouns throughout the whole input text, but rather switches pronouns, resulting in clear inconsistencies.

4.4 Verity

This chapter is closely related to culture-specific terms that can be found in internationally published text (DFKI, 2015). In this chapter, mistranslations of culture-specific terminology will be discussed. Overall, DeepL seemed to have very evident problems in translating culture-specific terminology and commonly used abbreviated forms of them.

(14) Error: Culture-specific reference

Original article by Ketonen, Petra (12.5.2022).

Pirte, Pirkanmaa [original Finnish text]
in Pirte, Pirkanmaa [DeepL initial translation]

In Example 14, a local healthcare service provider, Pirte, is discussed. Pirte is localized in Pirkanmaa, Tampere. DeepL mistakes Pirte for a region in Pirkanmaa, although the the word in question is a

name of this particular service provider. This example is particularly challenging for translating services since foreign names are easily mistaken for something else without the help of additional context. Gathering data to successfully recognize all the organizations in the world would take a vast amount of resources.

(15) Error: Culture-specific reference

Original article by Taleva, Katariina (16.5.2022, Iltalehti).

Tesyn Juup [original Finnish text]
Tesyn Juup [DeepL initial translation]

In Example 15, we can see another mistake in translating culture-specific references. The text mentions *Turun seudun eläinsuojeluyhdistys* [Turku Region Animal Welfare association], which DeepL is able to successfully translate. However, the abbreviation Tesy is not understood to represent the association, resulting in DeepL considering *Tesyn Juup* as a name. The correct translation for the sentence would have been Britt-Marie Juup from Tesy, or Juup from Tesy.

(16) Error: Culture-specific reference

Original article by Eskonen, Hanna. (18.5.2022)

Tehyn Rytkönen [original Finnish text]
Tehy Rytköne [DeepL initial translation]

Similarly to Example 15, a mistranslation occurs when the name of a culture-specific organization, in this case Tehy, The Union of Health and Social Care Professionals in Finland, is placed directly in front of the person representing the organization. DeepL is not able to recognize this relation, and treats the words as a name.

Overall DeepL does not sufficiently translate culture-specific terminology and abbreviated forms, which is a clear problem. but most likely very challenging to tackle. These translation models would have to be specifically trained to understand these culturally connected terms, which would take a vast amount of resources and new implementations. However, when it comes to news articles, they usually include local terminology, which can be unknown to the reader. Poor translations can therefore cause confusion and misunderstanding.

5. Conclusion

Although the analysis proves that DeepL is not a perfect translation service, I would like to further stress the points presented by Hutchins (2007). If MT manages to adequately translate to the point that the original text's core meaning is perfectly understandable to a reader familiar with the English language, I dare to argue that the translation is successful. If we target everyday conversations between people, or news articles written for the public, we do not need perfect translations, we need understanding. Also, I see that internet news outlets could definitely benefit from draft translations made from the original news texts. These draft translations would help the overwhelming task of manual translation to the point that news articles originally written in Finnish could be accessible for people not familiar with the Finnish language. This does not only apply to this specific language pair, as presented by the points by Hassan et al (2018, p.1), mentioned earlier in this study. However, if we look at the subject from a different perspective, as Cambedda et al. (2021) did when translating medical texts, adequate translations are not acceptable, and absolute perfection is demanded. Other fields that demand perfections are for example, legal translation.

The nature of this study being an error analysis, rather than a comparative analysis between two or more NMT services, makes it difficult to make claims about whether the previously listed claims made by DeepL are true. As Cambedda et al. (2021) showed in their comparative analysis between

DeepL, DeepL is able to outperform at least the NMT service Yandex. Still, it needs to be kept in mind that a different language pair, Italian-Russian, was used. Also, the scope of the study in question was to accurately translate medical documents, which take a different level of precision. The overall performance of a single NMT service could be incredibly dependent on the language pair used, considering the basics of how these services obtain their data. What is clear however, is that DeepL is an excellent translation service, but still far from perfect.

As presented by the analysis, DeepL is a very sufficient translation service regarding grammar. Only a few grammatical errors were detected throughout the analysis. Overall, the ability of DeepL to translate long sentences and change the word order to provide idiomatic translations, is impressive to say the least. The pronoun errors and the errors in consistency (multiple translations for one word) suggest that DeepL is still not sufficient enough in examining the resulting translation as a whole, but rather in smaller parts. The overall consistency of translated words in DeepL translations could be an interesting idea for future research.

Other clear flaws of DeepL include the translation of new words, abbreviations and symbolism. As presented in the analysis, most of the new words created during the COVID-19 pandemic were incorrectly translated, and were extremely inconsistent. It is however clear that gathering the corpora to accurately translate these words takes a lot of time and resources, and it would be interesting to examine how long does it take for DeepL to adequately translate an emerging new word in varying contexts. Abbreviations are another clear difficulty for DeepL, and for obvious reasons. Abbreviated terms are usually words that do not function alone without a definition, and they do not necessarily have a definition in the dictionary. Any set of words can be abbreviated, and if the translation engine does not have the non-abbreviated form as a part of the source text, or if the engine is otherwise not able to connect the term and the abbreviation, an effective translation is almost impossible to formulate. Symbolism in the case of this thesis was closely connected to idiomatic language use and the idioms of a language. If an uncommon word was used to symbolize something, DeepL translated the word literally. This phenomenon could definitely also be studied later.

I would also like to further elaborate the points made by Rivera-Trigueros (2021). The evaluation process done for the errors listed in this bachelor's thesis is not made by a certified professional, but rather by a student of the English language. In the case of this study, manual assessment was used solely on the basis of its simplicity, because generally used automated methods like BLEU

would have required a learning process which I did not see to be necessary for a study of this scale. Also, Rivera-Trigueros refers to the point originally made by Chatzikoumi (2020), which states that an evaluation process or study should preferably include two or more evaluators and calculate the inter-annotator agreement, but for this thesis, only one evaluator assessed all the errors collected for this study. Continuing with the statements made by Chatzikoumi (2020), I too think that the most suitable assessment method would be an implementation of both automated and manual methodologies, which in the case of this thesis would mean an addition of an automated method to analyze the resulted translations.

It is also to be noted that out of the 12 texts translated for the analysis of this thesis, only nine were analyzed further, since they had clear errors regarding intelligibility. Three of the translated texts did not present any mistakes that would give rise to confusion to a reader that we assumed to be familiar with English, also presuming that the focus point of "conveying the essence" is kept in mind.

I would also like to further emphasize that translation services like DeepL provide the user alternative examples for the translated text. These suggestions can be individually evaluated and the user can choose a word that best fits the context, this is, if they are familiar with the language used. Although this feature helps to develop these translation services, the originally given translation should naturally be as correct as possible. Unfortunately, it should be noted that due the suggested words may have the correct translation, the results of this study may not represent the full capability of DeepL, for that we assumed that the user is not familiar with L1. The accurateness of these given alterations would possibly be an interesting subject for study later.

Overall, the results of this study provide a concrete example of the current capability of NMT. The future of NMT seems promising, as new developments are applied frequently. NMT's ability to learn from both the user and from the collected data, makes this paradigm efficient to the point that I see no reason why translation systems could not be used to automatically translate news texts or other similar texts in the very near future.

List of references

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations*. Retrieved from ArXiv. https://arxiv.org/pdf/1409.0473.pdf

Cambedda, G. ., Di Nunzio, G. M., & Nosilia, V. (2021). A Study on Automatic Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation. *Umanistica Digitale*, (10), 139–163. https://doi.org/10.6092/issn.2532-8816/12631

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering. 26(2)*, 137-161. https://doi.org/10.1017/S1351324919000469

Corpus Linguistic Methods. (2014) What is Inter-Annotator Agreement? *Corpus Linguistic Methods. A Practical Introduction with R and Python*. Wordpress.

Cho, K., van Merrï¨enboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–Decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.*

DeepL. (November 1st, 2021) *How does DeepL work?* DeepL Translator. https://www.deepl.com/en/blog/how-does-deepl-work

Farrús, M., Costa-Jussà, M. R., Mariño, J. B., & Fonollosa, J. A. R. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. In *EAMT 2010—14th annual conference of the European Association for Machine Translation*. https://repositori.upf.edu/bitstream/handle/10230/34496/Farrus_EAMT2010_ling.pdf?sequence=1&isAllowed=y

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2015). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering, 23*(1), 3–30. https://doi.org/10.1017/S1351324915000339

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W.D., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X.,

Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., & Zhou, M. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. Retrieved from ArXiv. https://arxiv.org/abs/1803.05567

Hutchins, J. (2007) Machine translation: A concise history. *Mechanical Translation, 13(1 & 2) pp. 1-21.*

Kalchbrenner, N., & Blunsom, P. (2013, October 18). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP 2013)* (pp. 1700–1709), Seattle, Washington, USA.

Kim, Y., Graça, M., & Ney, H. (2020). When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020,* 35-44 https://arxiv.org/abs/2004.10581

Papineni, K., Roukos, S., Ward, T. & Zhu, W. (2002) *BLEU: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia. Pp. 311-318.

Rikters, Matīss. (2018) Impact of Corpora Quality on Neural Machine Translation. *Frontiers in Artificial Intelligence and Applications. 307,* 126-133. https://arxiv.org/abs/1810.08392

QT21 (2022) *Quality Metrics*. Quality Translation 2021 http://www.qt21.eu/quality-metrics/

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH / German Research Center for Artificial Intelligence (DFKI). (2015) *Multidimensional Quality Metrics (MQM Issue Types.* Retrieved from: http://www.qt21.eu/mqm-definition/issues-list-2015-05-27.html#accuracy

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: a systematic review. *Lang Resources & Evaluation* https://doi.org/10.1007/s10579-021-09537-5

Sutskever, I., Vinyals, O., & Le, Q. V. (2014, December 8-13). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27: Annual conference on neural information processing systems (NIPS)* (pp. 3104–3112). MIT Press

Way, A. (2018). Quality expectations of machine translation. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment* (pp. 159–178). Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_8.

List of translated texts

Eskonen, H., 2022. Kunnat ovat joutuneet maksamaan lääkäreille mitä hyvänsä, ja kohta sama tilanne voi koskea muitakin ammatteja – tutkija: ”Voi tarkoittaa käsittämätöntä kustannusten nousua”. Yle. https://yle.fi/uutiset/3-12442834

Hallamaa, T. (2022) Facebookin jakonappi levittää misinformaatiota, mutta suomalaiset painavat sitä harvoin. Yle. https://yle.fi/uutiset/3-12303545

Hannele, M. (2022) Työntekijöitä on sairaana ja karanteenissa, mutta pahin on vasta tulossa – epidemian huipulla sijaisia ei välttämättä löydy vuokratyöfirmoistakaan. Yle. https://yle.fi/uutiset/3-12272563

Juopperi, H., Annala, P. (2022) Oulussa jää keskiviikkona noin 250 bussivuoroa ajamatta – kaikista peruuntuneista vuoroista ei ole tullut edes tietoa Oulun joukkoliikenteen suunnittelijalle. Yle. https://yle.fi/uutiset/3-12267218?fbclid=IwAR1anho7pNSqaU1mt3mumlxSfqIQFJf29B01nu68Jte7_p0zjMcK7ov_Q88

Karttunen, A. (2022) Pronssimitalisti Iivo Niskasen käytös herätti hilpeyttä lehtistötilaisuudessa – toisti Christiano Ronaldon tempun. Yle. https://yle.fi/urheilu/3-12304560

Ketonen, P., 2022. Joillekin kertyy jopa 60 käyntiä vuodessa, osalle ei kuulu edes työterveystarkastuksia – tutkimus valottaa kiistellyn työterveyshuollon suurkulutusta. Yle. https://yle.fi/uutiset/3-12438325

Pikkarainen, A. (2022) Marin väläyttää nopeutettua rajoitusten purkua. Iltalehti. https://www.iltalehti.fi/politiikka/a/d817e44a-154e-4139-a6a2-4dc4cdc272c8

Puukka, P., Putkonen, J. (2022). Paola Suhonen puhui somevaikuttajista mutu-huutelijoina – nyt moni vaikuttaja ihmettelee, miksi yhteistyö on aiemmin Ivana Helsingille kelvannut. Yle. https://yle.fi/uutiset/3-12375965

Taleva, K., 2022. Korona-aikana hankituista koirista ja kaneista halutaan eroon – pelätty lemmikkien hylkäämisbuumi näyttää alkaneen. Iltalehti. https://www.iltalehti.fi/kotimaa/a/5d0e754a-ecf5-4a84-8202-d2184e78a709