Linguistic Error Analysis on Finnish to English Machine Translation

Anni Moisala
Bachelor's Thesis
English Philology
Faculty of Arts
University of Helsinki
May 2022

Table of Contents

List of Tables

List of Figures

1. Introduction

Today, more text is being translated by machines than human translators as a result of the ever-growing progress and availability of machine translation (MT). MT is transforming the field of translation, allowing fast and sufficient translations for personal use, companies, and even professional translators. To get an idea of the large scale in which MT is used, Google Translate alone hit 1 billion downloads in 2021 (Turovsky, 2021) and in 2016 it was translating 100 billion words a day (Pitman, 2016). The role of machine translation only keeps growing as the quality improves.

The biggest benefit of MT currently is its capability to translate almost instantly. Therefore, it can easily outperform humans in quantity of translation, yet it cannot reach the same quality as human translators. However, the ambition to attain human-like performance is clear in the research that is conducted to improve MT quality, especially in the field of error analysis.

This subfield of MT evaluation focuses on analyzing the errors made by MT systems in order to find how to improve these systems. The previous research includes studies of various methods and language pairs, such as automatic error analysis on Spanish output by Popović and Ney (2006), and comparisons of the output of different machine translation tools (Cambedda et al. 2021). Other studies have focused on creating error taxonomies, which are error categorization systems that allow systematic error analysis, such as the work by Costa et al. (2015) and Elliott et al. (2004).

I used a taxonomy by Costa et al. (2015) to perform error analysis on MT output from Finnish to English, to examine what types of errors occur with this language pair. The material consists of a parallel corpus that I compiled from news articles by the Finnish news broadcaster YLE and their official translations. Then I translated the Finnish articles to English using two different MT systems, Google Translate and DeepL Translator. Next, I annotated the translations for errors using the official translations as a reference, as well as my own judgement. My background is in English philology and translation studies, and I have native proficiency in Finnish. The results will indicate what needs to be improved to achieve successful translations with Finnish to English machine translation.

In the following section I will present the theoretical framework used in my study as well as give an overview of relevant previous research. In section 3, I give a detailed explanation of the materials and methodology used in this study. Results of the analysis will be presented in section 4, and further discussed and compared to previous studies in section 5. Finally, I conclude my findings and consider future studies.

2. Background

Evaluation of MT output is a fundamental part of the field of MT. It tells how functional a system is, how much it has improved and how to improve it further. Generally, evaluation can be divided into two main approaches: automatic evaluation and manual evaluation. Automatic evaluation uses different algorithms that calculate a score based on correspondence with a reference translation. These metrics include for example BLEU (Papineni et al., 2002), Meteor (Denkowski and Lavie, 2014) and chrF (Popović, 2015) evaluation scores. Manual evaluation, also known as human evaluation, can be done in many ways such as ranking of translation output, scoring translations, or doing detailed analysis.

Both methods have strengths and weaknesses and appropriate uses. Automatic evaluation is fast and cheap and easy to compare, but its biggest downside is being unreliable. Previous research has demonstrated that improvement in evaluation scores does not guarantee better translation quality (see e.g. Callison-Burch et al., 2006). Human evaluation is more reliable, it can provide detailed feedback and takes into consideration the fact that a sentence can have multiple different correct translations. However, human evaluation is time consuming and subjective. While automatic evaluation is good for getting instant feedback on the improvement of a translation model and comparing the performance of similar systems (Callison-Burch et al., 2006), human evaluation is needed to get reliable and detailed feedback, which in turn is essential for improving MT systems. As I would argue, a detailed linguistic analysis provides the greatest amount of information on how to improve the MT systems.

This linguistic analysis is generally performed by identifying and categorizing errors, also known as error analysis. Error analysis is a linguistic framework created to study the errors made by second language learners (James C, 1998). It has also proved to be very useful in manual evaluation of MT output since it aims to identify and analyze language mistakes with the goal of improving language use or in this case translation quality.

For this purpose, error taxonomies have been established to aid in categorizing errors, such as the work of Vilar et al. (2006), Litjós et al. (2005) and Elliott et al. (2004). These taxonomies focus on analyzing machine translated text, thus they include categories that consider errors specific to translation. For example, the taxonomy by Vilar et al. (2006) includes the following categories: missing words, word order, incorrect words, unknown words, and punctuation. This taxonomy builds up on the work by Litjós et al. (2005) and Elliott. et al. (2004), therefore these taxonomies share many of the error categories. These categories are further divided into subcategories to obtain more detailed

information. However, little attention has been paid to linguistic analysis of these errors. In contrast to the aforementioned taxonomies, this study uses a linguistically motivated error taxonomy by Costa et al. (2015). While previous taxonomies are generally built to serve the English language, this one aims to accommodate to more morphologically rich languages, i.e., languages with high inflection. With this in mind, they divided errors into five linguistic categories: orthography, lexis, grammar, semantics, and discourse. Costa et al. (2015) use their taxonomy to analyze errors in MT output of various systems from English to Portuguese, the latter being a morphologically rich language.

Considering the lack of related studies, linguistic analysis is underrepresented in the field of MT. However, to get a better idea of analyzing translation from a linguistic perspective, I will present some previous work on human translation analysis. First, the Multilingual eLearning in Language Engineering project (MeLLANGE), has established a learner translator corpus with annotated errors. For this they have created their own taxonomy, which divides errors into content related and language related errors, which are then divided into multiple subcategories (Castagnoli et al., 2007). In addition, Dulay et al. (1982) presents two error taxonomies: a linguistic category classification (LCC) and the surface structure taxonomy (SST). As described by their names, LCC is based on linguistic categories such as morphology, lexis, and grammar and SST focuses on omission, addition, misinformation, and misordering (Dulay et al., 1982 in Costa et al., 2015). Both of these approaches include detailed linguistic information, which MT evaluation taxonomies generally lack.

In addition to error taxonomies, MT output is often evaluated by two qualities: adequacy and fluency. These metrics are often judged manually by humans rather than by automatic evaluation systems. Adequacy measures how well the original meaning of the text is preserved, while fluency focuses on whether the text is well formed and written in a good and fluent manner (White et al., 1994). These metrics are a good way to judge the overall quality of a translation while providing some feedback on the properties of the text.

3. Materials and methods

In this section I describe the material and how it was collected, then give a brief overview of the two MT systems used in the study and lastly, present the methodology and explain the error taxonomy in detail.

3.1 Data

The materials for this study were compiled from news articles available freely online by YLE, Finland's national public broadcasting company. News articles provide language which is grammatically correct and neutral, and targets large numbers of people, making it suitable for this study. Three articles of roughly equal length were randomly selected from recent news at the time of conducting the study in 2022. The first article is about a collision of a cruise ship into a popular seaside pool in Helsinki, and I will refer to it as Text A. The second article describes the growth in popularity of the national parks in Finland, and I will refer to it as Text B. Lastly, the third article celebrates a Finnish skier Iivo Niskanen's bronze medal in the Olympic Skiathlon, and I will refer to it as Text C.

Only articles with English translations were selected, because these were used as a reference when performing the analysis. However, often these translations were shorter versions of the articles, limiting their use as a reference. These shorter versions included only the main points of the article. The Finnish articles were translated to English using Google Translate and DeepL Translator to create the MT output, which consists of roughly 1,800 words in total. Lastly, it should be noted that all the articles are published online for free and are available for the public, therefore there are no ethical issues concerning their use as a material for the study. In addition, the selected articles do not contain any personal information.

3.2 Translation systems

This study uses two state-of-the-art online machine translation systems: Google Translate and DeepL Translator. Google Translate is the most well-known and largest translation system, covering over a hundred languages and used by millions of people (Pitman, 2021). DeepL Translator is a little less known translation system launched in 2017, which covers 26 languages, however, it has proved to outperform many well-known translation systems (DeepL, 2020).

Both Google Translate and DeepL Translator are neural machine translation systems (NMT). This type of model, which is based on neural networks i.e., deep learning, entered the field of MT in

2014 and is the most dominant model today (Koehn, 2020). These neural networks, named after the neurons in a brain, consist of thousands of units that activate based on the stimuli received from other neurons and the connection along which they are passed. By simulating the human brain, although vaguely, neural networks allow the machine to learn from large amounts of data with little human intervention. This has benefited the field of MT greatly and it can be seen both in the popularity of NMT models and improvement in translation quality (Forcada, 2017).

3.3 Method of Analysis

My main approach to the linguistic error analysis is a qualitative close reading of the materials; however, it is also necessary to look at frequencies of errors in the data. The findings will be categorized based on an error taxonomy by Costa et al. (2015). This framework divides errors in five linguistic categories: orthography, lexis, grammar, semantic and discourse. These are then divided into multiple subcategories that describe the errors in more detail. However, only the first level of subcategories are included in my analysis and some of the subcategories with clear overlapping are merged. This way the level of detail is compatible with my dataset. Next, I give an overview of the different error categories in the order as listed above.

First, ORTHOGRAPHY level errors cover errors at the character level. This category is divided into PUNCTUATION, CAPITALIZATION, and SPELLING errors. PUNCTUATION errors represent misuse of punctuation, CAPITALIZATION errors occur when a word is unnecessarily capitalized or left uncapitalized, and finally SPELLING errors cover misspelling of words. However, CAPITALIZATION errors can be considered as SPELLING errors, thus these categories are merged in my analysis, and both are considered as the latter type.

Second, LEXIS level errors describe how and if the word was translated. These errors are divided into OMISSION, ADDITION AND UNTRANSLATED. OMISSION simply means that the word is missing, ADDITION stands for words that were not in the source text and UNTRANSLATED represents words that are left in the source language.

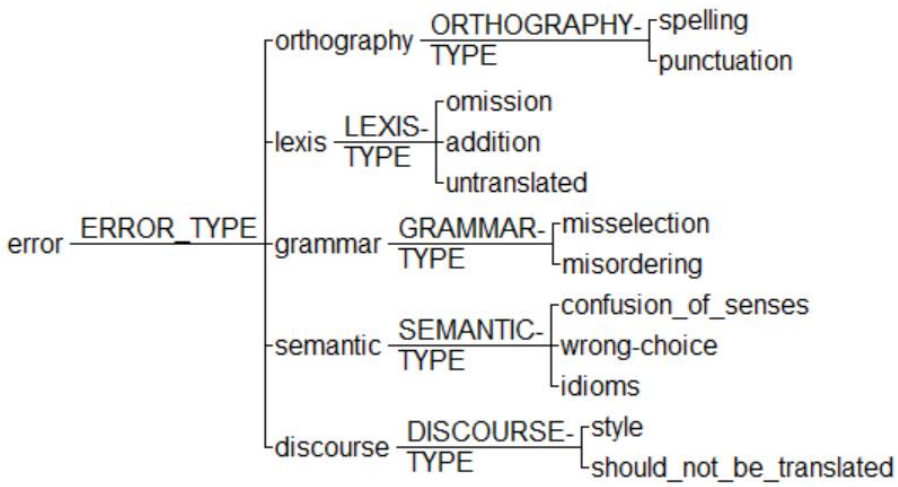
Next, GRAMMAR level contains two types of errors: MISSELECTION and MISORDERING. MISSELECTION errors happen when an incorrect word class is given, for example when a verb is needed but an adjective is given, tense and person is incorrect, gender, number or person does not agree or there is a contradiction between prepositions and articles. MISORDERING errors refer to wrong syntax.

SEMANTIC level covers errors that have to do with the meaning of the words, and it consist of four subcategories: CONFUSION OF SENSES, WRONG CHOICE, COLLOCATIONAL ERROR, and IDIOMS. CONFUSION OF SENSES happens when a word has multiple meanings and the wrong one is selected. In

contrast, when simply a wrong word is chosen, it goes under the category of WRONG CHOICE error. COLLOCATIONAL errors refer to mistranslations of expressions where two or more words go together. These can be considered as falling under CONFUSION OF SENSES errors. The distinction is made because CONFUSION OF SENSES errors cover one word and COLLOCATIONAL errors cover two or more words. However, this is not significant for my study, and I decided to merge them under the category CONFUSION OF SENSES. The category IDIOMS contains errors concerning the mistranslation of idiomatic expressions.

Lastly, DISCOURSE level errors contain errors that have to do with STYLE, VARIETY and WORDS THAT SHOULD NOT BE TRANSLATED. STYLE errors happen when a bad stylistic choice of words was made. VARIETY errors cover instances where the wrong variety of a language is used, but since Google Translate does not distinguish between varieties, I decided to not include VARIETY errors in my analysis. The last category, WORDS THAT SHOULD NOT BE TRANSLATED, covers words such as titles of movies, books, etc. that should be in the source language. The error categories and their subcategories included in my analysis are presented in Figure 1.

Figure 1: Error taxonomy based on Costal et al. (2015)



These error types were annotated in the data using an annotation tool called UAM Corpus Tool. The annotation is simply done by selecting a word from the text and then choosing the correct category. Statistics are then provided based on the annotation. The errors were annotated per error whether they covered one or more words. Thus, the results are presented as errors per dataset, and do not include word counts. It should also be mentioned that error analysis is not always straightforward: some errors occur across a sentence and are hard to pinpoint, overlapping of errors can occur and some cases can be subjective.

4. Results

In this section I present my findings starting with statistics and a brief comparison between the three texts that were analyzed, and two translation systems used to translate them. Then I present examples and further analysis of each error type.

4.1 Overview

First, I will present quantitative results to get an overview of how often the error types occur and in which texts. For simplicity, I have combined the different translation systems, but I will compare them separately in a different figure.

Table 1: Occurrences of each error type in the news articles

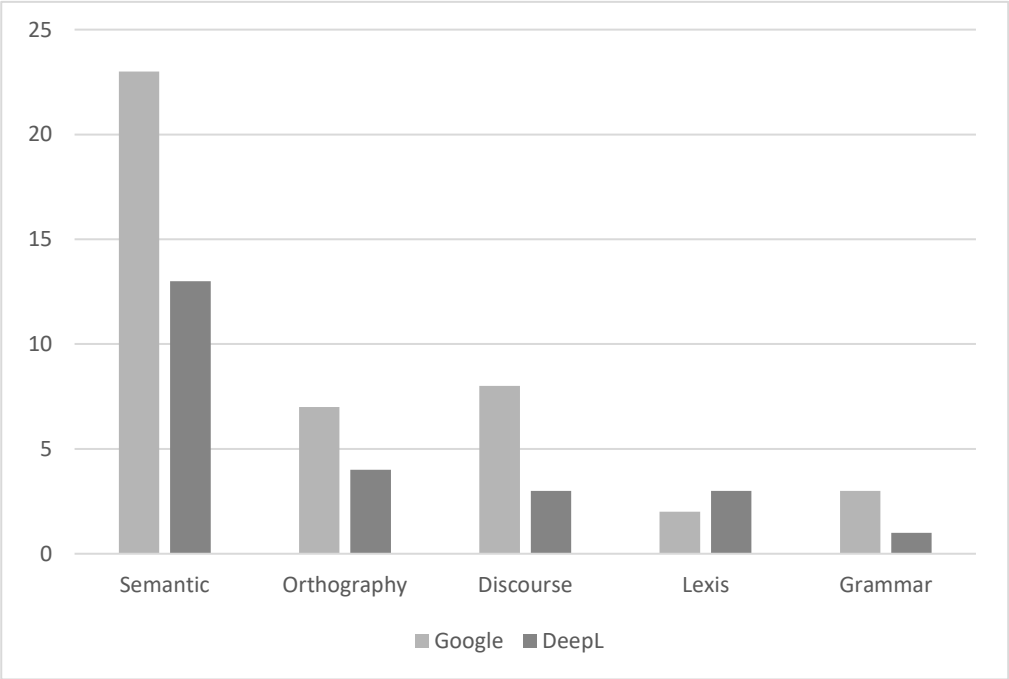
	Text A	Text B	Text C	Total
Semantic	11	5	20	36
Orthography	3	0	8	11
Discourse	5	6	0	11
Lexis	4	0	1	5
Grammar	1	1	2	4
Total	24	12	31	67

As can be seen from Table 1, the most common types of errors in the dataset are SEMANTIC level errors, which cover incorrect translations of the meanings of words, making up more than half, or 54% of all errors. Although the proportion is high, this is not surprising, since translating the meaning of a word is what translation at its core is about. The second place is shared between DISCOURSE level errors, consisting of stylistic errors and words that should not be translated, such as names, and ORTHOGRAPHY errors, concerning errors at the character level. Both types cover 16% of the errors. LEXIS level errors, which cover errors at the word level, such as omitting a word, make up 8% of the errors. The smallest category of errors in the dataset is GRAMMAR level errors, covering only 6% of all errors.

In addition, a comparison between the texts reveals that there are clear differences in the number of errors. Text A contains double the number of errors compared to Text B, and Text C almost three times as much as Text B. This can be explained by differences in the texts, for example Text C contains sentences from an interview, which were difficult to translate due to spoken language. Therefore, the text that is used for MT evaluation can have a notable effect on the outcome of the evaluation depending on how difficult it is to translate.

A comparison of the two MT systems shows a distinct difference in performance. As can be seen in Figure 2, DeepL Translator performed better in all categories, except lexis level errors. Even then, the difference between the two systems is only one error.

Figure 2: Comparison of the MT systems



4.2 Semantic level errors

The most frequent subcategory of errors in the SEMANTIC level is CONFUSION OF SENSES errors, which occurred 28 times in total. An example of that can be seen in Text A in the output of both MT systems in the translation of the word *sukellustutkimus,* meaning *diving investigation* or *examination*, where the word *survey* was used instead. While the word *tutkimus* can be translated as *investigation, examination,* or *survey* among many others, they have slightly different meanings. In this case, a more practical and hands-on word such as *investigation* is fitting, while *survey* refers to a list of questions aimed at studying something.

Errors belonging to the category of IDIOMS occurred in the dataset only four times, as idioms are

not very frequent in the texts. However, these types of errors are quite severe since mistranslations of idioms can produce full sentences that make little to no sense. As can be seen in Example (1), an idiom was translated literally in the output by Google Translate. In the examples, Acronyms GT and DL will be used for Google Translate and DeepL, respectively.

(1) Text C

FI: – *Pystyin pitämään homman hanskassa* loppuun saakka ja sain viimeisen kierroksen jopa vähän nautiskeltua, Niskanen myhäili Yle Urheilulle.

GT: - *I was able to keep the job in my glove* until the end and I got the last round even after a little enjoyment, Niskanen smiled at Yle Urheilu.

DL: - *I managed to keep it under control* until the end and even enjoyed the last lap a bit," Niskanen told Yle Urheil.

In this case, the translation of the idiom by DeepL Translator is correct. DeepL Translator proved to be more successful translating idioms than Google Translate in two other cases as well. Lastly, semantic errors categorized as WRONG CHOICE occurred four times in total. Hence, the MT systems are far more likely to select the wrong meaning for the specific context such as in CONFUSION OF SENSE ERRORS, rather than a wrong word altogether. First, as can be seen in Example (2), the word *korona* as in the *Coronavirus disease (COVID-19)* is mistranslated as *crown*. The correct translation would be *COVID-19 restrictions* or also known as *lockdowns.* However, even here the mistranslation is not completely random, and some relatedness exists.

(2) Text A

FI: Allas Sea Poolia ei voida avata yleisökäyttöön nykyisen *koronasulun* päättyessä helmikuun alussa.

GT: The Sea Pool cannot be opened to the public at the end of the current *crown closure* in early February.

DL: The Sea Pool will not be open to the public when the current *corona closure* ends at the beginning of February.

This error indicates that the data of the MT systems is not up to date, which can cause errors such as this. In comparison, DeepL translator translated it as *corona closure*, which classifies as CONFUSION OF SENSES error instead, due to semantic relatedness. While SEMANTIC errors cover the majority of errors in this dataset, there is no evidence in this case suggesting that they are specifically linked to qualities of the Finnish language or news articles, rather they are quite common in general.

4.3 Orthography level errors

ORTHOGRAPHY errors are not typical in MT because they rarely occur in the training data and therefore do not appear in the translations. However, in this dataset, ORTHOGRAPHY errors are the second most common with 11 errors overall. Most of these, nine in total, are PUNCTUATION errors. For example, in Text A Google Translate output, the name of a ship, M/S Gabriella, is spelled incorrectly with spaces, as M / S Gabriella. This was the case with all three occurrences of the name.

The remaining six PUNCTUATION errors were caused by quotations: the quotations in the source text are marked with a dash, which is common practice in Finnish but not used in the English language. Therefore, the translation systems fail to change the quotation style, although quotation symbols are added at the end of the quote, as can be seen in Example (2). Furthermore, the dash, used for quotations, is replaced with a hyphen. In addition, the example contains a SPELLING error in the name of the sport news broadcaster Yle Urheilu.

(2) Text C

FI: – Pystyin pitämään homman hanskassa loppuun saakka ja sain viimeisen kierroksen jopa vähän nautiskeltua, Niskanen myhäili *Yle Urheilulle.*

DL: - I managed to keep it under control until the end and even enjoyed the last lap a bit," Niskanen told *Yle Urheil.*

In the Finnish text, the name *Yle Urheilu* is inflected to *Yle Urheilulle*, to express the dative case, in other words, to indicate the recipient of the action. The misspelling occurs when the inflected form of name needs to be converted to the uninflected form for the English translation. A similar mistake is made in the same text when the inflected form of the name *Niskanen* is incorrectly spelled as *Niska*, as shown in Example (4).

(4) Text C

FI: Norjalainen Hans Christer Holdund näytti aluksi saavuttavan *Niskasta*, mutta jäi lopulta puoli minuuttia suomalaisesta.

GT: The Norwegian Hans Christer Holdund seemed to reach *Niska* at first, but ended up half a minute behind the Finn.

Instead of typos in the training data, which would usually cause SPELLING errors in the MT output, these errors were caused by words, more specifically names, that were likely not included in the training data of the MT systems. Generally, these unknown words cause semantic errors, but these examples showcase a challenge that is unique to languages with high inflection, such as Finnish. Furthermore, since both instances were caused by possibly unknown words, it suggests that known words do not cause spelling errors due to inflection. However, these types of errors were quite rare, as there were only two instances of them. It would be interesting to see how often and when they would occur in a bigger dataset.

4.4 Discourse level errors

In the category of DISCOURSE errors, six or about half of them fall under the subcategory STYLE errors. These are the most subjective type of error and identifying them is not as straightforward as with other error types. However, one common style error MT systems make is repeating a word inside a sentence. Here is one such case in Example (5), where the word *visit* occurs twice in the sentence by Google Translate.

(5) Text B

FI: Esimerkiksi Repoveden kansallispuistossa Kouvolassa *vieraili* viime vuonna noin 173 000 *kävijää*.

GT: For example, the Repovesi National Park in Kouvola was *visited* by about 173,000 *visitors* last year.

DL: For example, Repovesi National Park in Kouvola *had* around 173 000 *visitors* last year.

The sentence by DeepL Translator is more pleasant to read compared to the one by Google Translate, which contains repetition. Another stylistic error in Text B is the use of the word *teleworking*, which is a correct translation for *etätyö.* However, a much more preferred term during COVID-19 has been *remote working* or *working from home*. A Google search reveals that *remote working* returns 146 times more search results and *working from home* 478 times more search results than *teleworking*, which is a significant difference. While the error is not severe, it is interesting to note how slowly MT systems adapt to new terms, and thus can provide outdated translations.

The second subcategory of DISCOURSE errors is words that SHOULD NOT BE TRANSLATED, and these occurred five times in the dataset. Typically, these include names such as titles of books and films. Similarly, all occurrences in this dataset were caused by the name *Allas Sea Pool,* which is quite unique because it is a combination of Finnish and English, providing a challenge for MT systems. Naturally, in many cases the word *Allas* was translated to *pool*, as can be seen in Example (6). However, not all cases were mistranslated, showing that it is possible for the MT systems to detect it as a name.

(6) Text A

FI: Kauppatorin kupeessa sijaitseva *Allas Sea Pool* on erityisesti turistien suosima kohde.

GT: Located next to the market square, the *Pool Sea Pool* is especially popular with tourists.

4.5 Lexis level errors

With LEXIS level errors, we are again concerned with the name *Allas Sea Pool.* In addition to erroneously translating the Finnish word *Allas*, in some cases it was left out, falling under the category of OMISSION, as can be seen in Example (7):

(7) Text A

FI: *Allas Sea Pool* sai osumaa jättimäisestä risteilijäaluksesta toissa viikolla ja pysyy suljettuna, kunnes vahinkojen kartoitus on valmis.

DL: *Sea Pool* was hit by a giant cruise ship the week before last and will remain closed until a damage survey is completed.

In most cases, Google Translate would translate the world *Allas*, while DeepL Translator omitted it. Generally, OMISSION errors were quite rare, occurring in only two cases. The previous error was repeated four times, bloating the number of lexis errors. Finally, the fifth OMISSION error is the omission of an adjective that precedes the word victory, shown in Example (8):

(8) Text C

FI: Kilpailun *ylivoimaiseen voittoon* hiihti venäläinen Aleksandr Bolshunov.

GT: Aleksandr Bolshunov from Russia skied to the *victory* of the competition.

DL: Aleksandr Bolshunov from Russia *won the race by a landslide*.

The omitted word *ylivoimainen* can be translated as *superior, overpowering, or outstanding*. Then why was it left out? It is unlikely that the Google Translate does not know the word and thus would leave it out. However, it seems that in this case, simply translating the word does not produce a fluent sentence. Consider the sentence *Aleksandr Bolshunov from Russia skied to superior victory of the competition.* The word *superior* does not seem to work here and the OMISSION error in this case produces a more fluent sentence. However, as can be seen in the sentence by DeepL Translator, MT is capable of better translation without omitting anything. Lastly, regarding the subcategories ADDITION and UNTRANSLATED, these error types did not occur in the dataset. This result does not surprise, since these types of errors are rare with high quality MT.

4.6 Grammar level errors

Overall, the grammar of the translations produced by the MT systems was of high quality, and grammar level errors occurred only four times, being the rarest error type in this dataset. Surprisingly, the differences in grammar between Finnish and English did not cause many grammar errors. That being said, some such errors occurred, as can be seen in Example (9), where the verb *suffer* is in the wrong tense, causing a MISSELECTION error.

(9) Text A

FI: Laiturirakenteet, saunarakenteet ja meren puoleiset kulkureitit *kärsivät* eniten.

GT: Pier structures, sauna structures and seaward routes *suffer* the most.

DL: The dock structures, sauna structures and sea-facing walkways *suffered* the most damage.

This can be explained by the Finnish word *kärsivät* being both the past and present tense of the third person plural form. The example shows how the inflections of a synthetic language can cause mistranslations. However, DeepL Translator translated the word correctly. The Finnish word order is also quite different from the English one, therefore WORD ORDER errors may occur. Word order is typically more important in analytical languages such as English. In the Example (10), the Finnish word order is kept in the English translation, which is not a typical word order in English.

(10) Text B

FI: Retkikohteista suosituin on Pallas-Yllästunturin kansallispuisto, jossa kävi *viime vuonna* miltei 700 000 kävijää.

DL: The most popular destination is the Pallas-Yllästunturi National Park, which *last year* attracted almost 700 000 visitors.

GT: The most popular excursion destination is Pallas-Yllästunturi National Park, which was visited by almost 700,000 people *last year*.

In English, adverbs such as *last year* are typically placed at the end of the sentence, as in the translation by Google Translate. In conclusion, while Finnish and English have very different grammar, these types of errors are still quite minimal. This shows that the translation systems are able to learn the English grammar quite well, and mainly produce text based on these rules, and grammar errors are avoided as long as the meaning of the source text was grasped.

5. Discussion

A quantitative analysis of my results shows how the errors rank in terms of quantity. Next, I will evaluate their effect on translation quality based on previous research. First, SEMANTIC level errors were the most common, showing that MT systems struggle the most with picking the correct meaning of words. In addition, research by Costa et al. (2015) shows that SEMANTIC errors have the biggest effect on translation quality. Therefore, improvement on the SEMANTIC level would be the most beneficial of all these error types for the quality of MT. Moreover, the adequacy of a translation, describing how well the meaning of the original text is preserved in the translation, has been shown by studies to rank lower in evaluation scores compared to fluency (Martindale et al., 2019). My study supports these findings since SEMANTIC level errors affect the adequacy of a translation rather than its fluency, and high quantities of SEMANTIC level errors cause low rankings in adequacy scores.

Second, ORTHOGRAPHY errors, covering spelling errors, and DISCOURSE level errors such as STYLE errors, share the place of second most common error types in my study. In contrast to SEMANTIC level errors, these are less severe errors, having the lowest effect on translation quality according to Costa et al. (2015). In addition, these error types primarily affect the fluency of a translation, which correlates to the consensus that fluency in MT generally ranks higher than adequacy. However, while research indicates that these type of errors have a lesser impact on translation quality, my research points out that they rank high in quantity in Finnish to English machine translation.

Lastly, in my results, GRAMMAR errors and LEXIS level errors such as OMISSION rank the lowest in quantity. In severity, these rank lower than SEMANTIC level errors but higher than ORTHOGRAPHY and DISCOURSE level errors, according to Costa et al. (2015). Interestingly, these also fall in between affecting the adequacy and fluency of a translation. Therefore, errors that have to do with GRAMMAR and LEXIS do not seem to be the most urgent problems in MT quality.

Due to lack of previous research on error analysis of MT output using linguistic categories, it is not possible to draw conclusions on how the language pair Finnish-English affected the ranking of these error types. It should also be mentioned that the nature of my study poses some limitations. First, the size of the material is not sufficient to draw substantial conclusions about the ranking of the error types. Although the results show a general distribution of the errors, these results may look different on a larger dataset. However, the size of the material allowed me to carefully annotate each text and analyse the errors. This supports my goal of detailed linguistic analysis of machine translation errors, where results of a qualitative analysis are the main concern.

Second, limitations concerning my method should also be considered. Since the annotation is

performed only by one person it is more subject to mistakes and subjectivity. Ideally, there would be multiple annotators to achieve more objective results and catch possible mistakes. However, I simulated this by comparing my results to the official translations, and the annotation was done with academic background in English philology and translation studies, and native proficiency in Finnish. Therefore, these limitations should not severely affect the validity of my results.

The qualitative analysis revealed how the language pair Finnish-English affects the errors that occur. Some error types where found to be more dependent on the language pair than others. Moreover, error types that are less affected by the properties of the languages cover SEMANTIC, DISCOURSE and LEXIS type errors, while error types often caused differences in morphological structure of the languages cover ORTHOGRAPHY and GRAMMAR level errors.

Errors independent of the language pair were caused for example by mistranslations of the meaning of words, i.e. SEMANTIC errors. Other errors include stylistic errors, falling under the category of DISCOURSE errors, that were caused by the vocabulary of the MT system not being up to date. Lastly, LEXIS level errors such as translating words that should not be translated, can be linked with MT systems not recognising names that should not be translated.

Errors dependent of the language pair in this study cover ORTHOGRAPHY and GRAMMAR level errors. Since Finnish is a morphologically rich language, i.e. highly inflected, this causes the majority of the language dependent errors. This can be seen when translating the inflected forms of Finnish names into uninflected forms for the English translation causes incorrect spelling of the words, specifically with names (see Examples 4 and 5). Another example included a GRAMMAR error caused by a different word order convention (see Example 9).

As machine translation is constantly improving, it begs the question whether it is possible to achieve flawless machine translation and if not, what is the end goal. While some argue that MT has almost reached human parity, others say it is not possible. Yet for some purposes it is already fitting, such as translating webpages and even speeding up the work of human translators. Although MT might not be fit for every translation task, it still has an important role. For example, it aims to advance fair access to information by removing language barriers in situation with groups such as refugees, migrants, and people in crisis situations (Nurminen and Koponen 2020). Machine translation is most needed in situations where human translation is not possible, whether it is because of accessibility, funds, or time.

6. Conclusion

This study has analyzed error types in translations by two MT systems Google Translate and DeepL Translator of news articles from Finnish to English. The errors were annotated according to a linguistically motivated taxonomy by Costa et al. (2015). Quantitative results of the annotation showed that semantic level errors covered more than half of the errors, orthography and discourse level errors ranked second and lastly lexis and grammar level errors only made up a small portion of total errors. In addition, I analyzed the effect of the language pair on the errors. The study found that features of the Finnish language were detected as the cause of some errors, mainly orthography and grammar level errors. The biggest cause was the inflected forms in Finnish, which caused misspellings of words when they needed to be converted into an uninflected form for the English translation. Since these types of errors did not occur as often in the data, a study with a bigger sample would provide more information about how successfully MT systems can convert inflected forms back to their uninflected form.

As machine translation has seen huge progress with the introduction of neural networks, and the use of machine translation keeps increasing among individuals, companies, and professional translators, providing high quality machine translation has become increasingly important. With the near human parity that machine translation has reached, it has also become necessary to perform detailed error analysis to see how these systems have improved and what needs to be improved in the future. This study has pointed out the types of errors state-of-the-art NMT models do on translations from Finnish to English, and the results indicate a research direction for how these MT systems need to be improved.

References

Primary sources and translation systems

DeepL Translator, DeepL SE, Cologne, Germany. Available at: https://www.deepl.com/translator [Accessed 27 April 2022]

Google Translate, Google. Available at: https://translate.google.com [Accessed 27 April 2022]

Karhu, O., 2022. Text A: Allas Sea Pool sai osumaa jättimäisestä risteilijäaluksesta toissa viikolla ja pysyy suljettuna, kunnes vahinkojen kartoitus on valmis. *Yle Uutiset*. Available at: https://yle.fi/uutiset/3-12291698. Engl: Popular Helsinki seaside pool damaged in Viking Line's recent pier collision. Available at: https://yle.fi/news/3-12292092 [Accessed 4 April 2022]

Korpela, H., 2022. Text B: Kansallispuistojen suosio jatkui viime vuonna: käyntimäärät ylittivät ensimmäisen kerran neljän miljoonan rajan. *Yle Uutiset*. Available at: https://yle.fi/uutiset/3-12291211 Engl: Record numbers visit Finland's national parks in 2021. Available at: https://yle.fi/uutiset/3-12291211 [Accessed 4 April 2022]

Saarinen, O., 2022. Text C: Iivo Niskanen avasi Suomen mitalitilin olympialaisissa! Komea pronssimitali yhdistelmähiihdosta. *Yle Uutiset.* Available at: https://yle.fi/urheilu/3-12304433. Engl: Niskanen takes bronze in Olympic Skiathlon. Available at: https://yle.fi/news/3-12304463 [Accessed 4 April 2022]

Secondary sources

Callison-Burch, C., Osborne, M. and Koehn, P., 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In: EACL 2006*, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference.* Trento, Italy, April 3–7, 2006*.* Association for Computational Linguistics, pp. 249–256. Available at: https://www.aclweb.org/anthology/E/E06/E06-1032.pdf [Accessed 4 April 2022]

Cambedda, G., Di Nunzio, G. M. and 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), pp. 139–163. DOI: https://doi.org/10.6092/issn.2532-8816/12631

Castagnoli S., Ciobanu D., Kunz K., Volanschi A. and Kubler N., 2007. Designing a learner translator corpus for training purposes. In: TALC7, *Proceedings of the 7th teaching and language corpora conference*. Paris, France. Available at: https://www.academia.edu/3801199/Designing_a_Learner_Translator_Corpus_for_Training_Purposes [Accessed 4 April 2022]

Costa, Â., Ling, W., Luís, T., Correia, R. and Coheur, L., 2015. A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, 29(2), pp. 127–161. DOI: https://doi.org/10.1007/s10590-015-9169-0

DeepL Translator, 2020. *Another breakthrough in AI translation quality*. Available at: https://www.deepl.com/en/blog/20200206 [Accessed 28 April 2022]

Denkowski, M. and Lavie, A., 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation.* Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics, pp. 376–380. DOI: 10.3115/v1/W14-3348

Elliott D., Hartley A. and Atwell, E., 2004. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. *Springer*, pp. 64–73. DOI: https://doi.org/10.1007/978-3-540-30194-3_8

Forcada, M. L., 2017. Making sense of neural machine translation. *Translation Spaces*, 6(2), pp. 291–309. DOI: 10.1075/ts.6.2.06for

James, C., 1998. Errors in language learning and use. Exploring error analysis, applied linguistics and language study. *Routledge, New York*.

Koehn, P., 2020. Neural machine translation. First edition. *New York: Cambridge University Press,* pp. 39–40. DOI: https://doi.org/10.1017/9781108608480

Litjós, A. F., Carbonell, J. G., Lavie, A., 2005 A framework for interactive and automatic refinement of transfer-based machine translation. In: 10th EAMT conference *Practical applications of machine translation*. Budapest, Hungary, May 30–31, 2005. European Association for Machine Translation, pp 87–96. Available at: https://aclanthology.org/2005.eamt-1.13 [Accessed 27 April 2022]

Martindale, M., Carpuat, M., Duh, K., McNamee, P., 2019. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. In: *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin, Ireland, August 2019. European Association for Machine Translation, pp 233–243. Available at https://aclanthology.org/W19-6623 [Accessed 27 April 2022]

Nurminen, M. and Koponen, M., 2020. Machine translation and fair access to information. *Translation Spaces*, 9(2), pp. 150–169. DOI: https://doi.org/10.1075/ts.00025.nur

Papineni, K., Roukos, S., Ward T., and Zhu W., 2002. BLEU: a method for automatic evaluation of machine translation. In: ACL 02, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, July 7–12, 2002. Association for Computational Linguistics, pp. 311–318. DOI: https://doi.org/10.3115/1073083.1073135

Pitman, J. 2021. Google Translate: One billion installs, one billion stories. Available at: https://blog.google/products/translate/one-billion-installs/ [Accessed 4 April 2022]

Popović, M. and Ney, H., 2006. Error analysis of verb inflections in Spanish translation output. In: *TC-STAR workshop on speech-to-speech translation.* Barcelona, Spain, June 2006. TC-Star, pp 99–103. Available at: https://www.researchgate.net/publication/228945619_Word_error_rates_Decomposition_over_POS_classes_and_applications_for_error_analysis [Accessed 4 April 2022]

Popović, M., 2015. chrF: character n-gram F-score for automatic MT evaluation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation.* Lisbon, Portugal, September 2015. Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049

Turovsky, B. 2016. Ten years of Google Translate. Available at: https://blog.google/products/translate/ten-years-of-google-translate [Accessed 10 April 2022]

Vilar, D., Xu, J., D'Haro, L.F. and Ney, H., 2006. Error analysis of machine translation output. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy, May 2006. European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf [Accessed 4 April 2022]

White, J. S., O'Connell, T., O'Mara, F., 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, Maryland, USA, October 5–8, 1994. Available at: https://aclanthology.org/1994.amta-1.25/ [Accessed 29 April 2022]