TURUN YLIOPISTO

Neural Machine Translation and Finnish Case-Inflections: Translation Problems and Pre-editing Possibilities

Mikael Rantanen

Master's Thesis

Degree Programme in Multilingual Translation Studies, Department of English
School of Languages and Translation studies
Faculty of Humanities
University of Turku

March 2024

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -järjestelmällä

Master's Thesis

Degree Programme in Multilingual Translation Studies, Department of English
Mikael Rantanen
Neural Machine Translation and Finnish Case-Inflections: Translation Problems and Pre-editing possibilities
Number of pages: 50 pages, 15 pages in appendices

Machine translation has slowly been becoming more prevalent from the 1950s onwards with the most recent leap in development being neural machine translation, which has quickly been deployed in professional translation settings with human translators post-editing machine translated text. Neural machine translation is still not a perfect form of machine translation. One of the problems when translating between English and Finnish is the Finnish case system, as the Finnish case-inflections can be problematic for machine translation engines.

This thesis presents an evaluation of the four most prominent neural machine translation engines' capability of translating case-inflections accurately from English to Finnish and proposes pre-editing or pre-translation being a solution for the problem. The four neural machine translation engines that are featured are Google translate, Microsoft translator, Amazon translate, and DeepL translator.

Samples were gathered from news media, a fictional book, and a scientific book to account for differences in textual style. These samples were then given as input to neural machine translation engines and the samples' translations were given a numerical score on a seven-levelled evaluation scale. The samples that were assigned scores which were deemed unsatisfactory were selected and the possible reasons for their inaccuracy was examined.

The further examination showed that the frequency in which a case-inflection occurs predicts it's accurate usage by a machine translator to some degree. Unconventional syntax structure and longer sentence length were also shown to be predictors for accuracy problems when translating with a machine translator. The results suggested that pre-editing would not be more efficient than post-editing.

Key Words: machine translation, neural machine translation, case-inflection, pre-editing

Table of contents

1 Introduction

For centuries, people have had the need to communicate with other people, many of whom speak different languages. The need for cross-language communication has been growing at the same pace as the world's population has grown. The need grew more rapidly than ever before after the industrial revolution, when the worldwide trade escalated. With the introduction of information technology, mainly through the invention of the personal computer and the world wide web, people from everywhere on the globe have become able to communicate with each other as easily as pressing a few buttons to write an email or a text message or pressing only one button to send a voice message or calling on the phone. The phenomenon that we now call globalization has introduced an unprecedented need to communicate quickly and accurately with people who do not share the same mother tongue than oneself. Thus, the need for translators and interpreters has risen. The English language has quietly established itself as a sort of lingua franca in the globalized world. In many countries, English is taught in schools as a secondary or tertiary language, and it is increasingly frequent that one can get by anywhere in the world by only speaking English. This phenomenon has, in its own right, decreased the need for translations in everyday life. In official contexts, translations must still be performed by trained professionals to make sure that people can understand the texts.

For decades, researchers in the field of language have sought for solutions to the rising need for translation services. The first known translating machine, which the inventor dubbed the "Mechanical brain", was patented in 1933 by French-Armenian engineer George Artsruni (Henisz-Dostert et al. 1979, 7). At the same time, a USSR-based researcher P. P. Trojanskij theorized a detailed process of translating text from one language to another. These have been thought to be the earliest models of automated translation but did not yet set in motion the large-scale research of machine translation (ibid.). In January 1954, IBM introduced a computer capable of translating Russian into English (Angelone, Ehrensberger-Dow, and Massey 2020, 285). This has been thought of as the moment translation technology surfaced as a sub-domain of information technology, although, after its initial conception it was referred to as human language technology (HLT) instead of translation technology (ibid.).

Machine translation is ever evolving and in the eight decades the humans have had the notion of machine translation, multiple different models of MT have been conceived. The most prominent models include rule-based machine translation (RBMT), statistical machine

translation (SMT), and neural machine translation (NMT). The earliest MTs used rule-based methods. In 1988, IBMs Peter Brown suggested that SMTs could be better translation engines than RBMTs that were mostly used until then, and many scholars and other major figures in the field of translation technology did, at first, not agree to this sentiment (Angelone, Ehrensberger-Dow, and Massey 2020, 313). Despite the early opposition, by the mid 1990s the SMT had become the more prominent MT model (Angelone, Ehrensberger-Dow, and Massey 2020, 315). SMTs were further improved in 2007 by implementing a phrase-based SMT model (PBMT), which became the most prominent model for the following 10 years until the first NMTs were starting to be developed in the mid 2010s (Angelone, Ehrensberger-Dow, and Massey 2020, 316). The first NMT models did not perform as well as PBMT models, but improvements were quick with the implementation of an encoder–decoder model and an attention model which combine to make an encoder–decoder model with attention, which is prominently used to this day (ibid.). According to Koehn (2020, 31) the first neural networks were conceived as early as the late 1950s by Rosenblatt (1957). This neural network model consisted of only a single layer but was able to learn basic mathematical operations (ibid.). The first multiple layer neural networks were proposed in the 1960s, but they did not acquire much interest until the mid-1980s (Koehn 2020, 32).

This thesis focuses on four different NMTs: Amazon translate, Google translate, Microsoft translator, and DeepL translate, and their ability to translate linguistic cases from English to Finnish. All of the MT engines chosen for this research are based on neural networks.

Earlier machine translation models, such as SMTs and RBMTs, are currently outshined in many areas by the newer NMT models, at least in terms of translation quality (Nitzke, and Hansen-Schirra 2021, 26). Modern NMT models can be useful in domains where an MT has not been usable before, such as literary and audiovisual translation (ibid.). The main problem with the NMT models is that they require massive amounts of high-quality training data to be useful at all (ibid.).

The Finnish language features a case system with 14 different case-inflections in addition to the nominative. Seven of the 15 case-inflections will be the target of this study. As presented in Jaakola's and Onikki-Rantajääskö's book *The Finnish Case System: Cognitive Linguistic* Perspectives, the 15 case-inflections that exist in the Finnish language correspond to 19 constructs in English and are usually manifested as prepositions (Huumo 2023, 18). This thesis tackles the problems that NMTs have in translating textual entities containing one or

more the aforementioned seven prepositions into Finnish. The aim is that the translated text contains at least one of the seven case suffixes that are prevalent in the Finnish language. The translation into case suffixes could be made to always be correct by a simple rule-based method, such as: “if X then Y”. This should be an easy parameter for a machine translation engine to follow. The two most prominent problems seem to be, firstly, that the English language structures cannot be predicted to directly correspond to their 14 counterparts in the Finnish language, not including the nominative, and secondly, the context must be interpreted correctly for any rule to be properly implemented. This would result in a complex network of parameters that would be impossible to even construct.

Finnish case-inflections occur in varying frequencies within most language use. The frequency that each case-inflection occurs in typical language use can be predicted to be reflected in an NMTs training data, which means that the more frequently occurring case-inflections are more likely to be used correctly by an NMT and the less frequently occurring case-inflections to be less likely to be used correctly. The other possibility is that the NMT works around the use of the less frequently occurring case-inflection in favor of other structures that are more frequently used in the training data.

The broader scope that this analysis of the NMT engines’ capability enlightens is the argument that NMTs require a vast amount of learning data before they can properly operate compared to SMTs and RBMTs, which work quite well with little data. This results in hugely unequal results between different languages, as there is a considerably bigger amount of data available for, for example, the tenth most spoken language in the world than the 100th most spoken language. The problem with the amount of training data can also exist within a language in the case of words that rarely appear, especially in written language. The text samples used in this study will be chosen in such a way that the case-inflections appear in their most common form, that is, modifying a noun and fulfilling their most typical purpose. The Finnish case-inflections do have specialist use-cases, but they were intentionally not included for the sake of consistency.

Today, post-editing (PoE) has largely replaced traditional translation, at least between languages, for which, large language repositories, such as corpora can be found. Post-editing is defined as “the correction of raw machine-translated output by a human translator according to specific guidelines and quality criteria” (O’Brien in Nitzke and Hansen-Schirra 2021, 8). The text that is to be post-edited may have been divided into segments or appear as a

whole text. The division into segments is most prominently the case when using a computer assisted translation (CAT) tool such as Trados Studio or Phrase TMS. The first commercial translation memories were released in the 1990s and from then on post-editing has become increasingly more used in the realm of professional translation. Before this, pre-editing (PrE) was prominent in some contexts because RBMTs and SMTs could be effectively used in specific domains. Pre-editing, broadly defined, is a procedure, where a text is edited to match a predetermined ruleset before it is inputted to the MT engine. Pre-editing traditionally involves a pre-determined ruleset that involves the use of specific grammar, syntax, and lexicon, according to which the text should be pre-edited. Pre-editing has previously mostly been used to edit technical texts and text types that are typically already strict in grammatical and syntactical rules and feature a strict terminology. Pre-editing has not been extensively research using NMTs and the research has traditionally been focused on pre-editing of the output of SMTs. Pre-editing does not seem to be useful enough for texts that are written with less strict grammatical and syntactic structure and that feature a broader lexicon, such as fictional works. Nevertheless, pre-editing of portions, whose translation can be predicted to feature a case-inflection could result in a more accurate MT output that does not need to be post-edited as much, if at all. In this study, I will consider the possibility that pre-editing could be useful for a specialist use-case, more specifically getting an MT to translate case-inflections accurately without, or with little, further human intervention.

The translation samples that are used as the dataset for this thesis will be assigned a numerical score according to the accuracy of the translation of the predicted case-inflection. Because a complex issue such as correct case-inflection usage cannot be accurately determined by a numerical value, I will examine the correctness from the perspective of a human translator and suggest possible reasons for the possible issues considering the correlation to the frequency of occurrence of the case-inflection and contrasting the findings to findings from previous research and other theoretical background material. After examining the findings with from a qualitative perspective, I will consider solutions to the possible problems that the NMTs encountered during the data gathering phase and the possible tribulations that might be encountered when implementing the suggested solutions with the support from the theoretical background material.

2 Theoretical background

The study that is produced in this thesis is, at least to the author's knowledge, the first one recorded with the specific setting and, without a doubt, the first that has been produced with the specific dataset. This means that there is little to no prior research to draw from. Nonetheless, studies providing supporting results have been conducted prior to conducting the study presented in this thesis. I will draw supporting and contesting arguments from earlier research that has been conducted in the field of neural machine translation.

2.1 Neural Machine Translation

2.1.1 NMTs as the new norm

Nowadays, NMTs are highly regarded in the translation field. In the translation industry of today, it is more infrequent to translate documentation without the help of an MT and a TM, especially if one or both of the source language and the target language are ones that are among the top 10 most used languages in the world. One exception to this are fictional texts, whose translation requires translation i.e. translating a text, while simultaneously creating content for the translation. Post editing has become the new norm for translators. In his 2020 book, Philipp Koehn explains neural networks in detail. Below is a brief overview of the operation of neural networks in the linguistic context based on Koehn's explanation.

Firstly, a simpler linear model used in traditional statistical models: when the linear model is given a sentence to translate, it will use a set of weighted parameters to decide, which features of the target language it will use to produce the translation (Koehn 2020, 67). This means that the engine has been 'trained' in a way that assigns weight to linguistic units, for example: words or morphemes, which can then be used by the engine in the translation (ibid.)

The NMT uses a basis of this linear function, but also introduces a set of hidden values, which are weighted instead of the output values (Koehn 2020, 68). One of the more important distinctions at this moment is that there are matrices of weights connecting the input nodes and the hidden nodes, as well as the hidden nodes and the output nodes (ibid.). After these operations, the function is run through a nonlinear function (Koehn 2020, 69). This nonlinear function can result in scenarios where the node is turned off completely, the node is partly turned on, or the node is completely turned on (Koehn 2020, 70). These layers are stacked on

top of each other (Koehn 2020, 68). The more layers the more probable it is that the output is of high quality (Koehn 2020, 70).

In 2016, Google, who had released their original Google translator that was based on a SMT model, became the first company to release a commercial NMT system (Nitzke and Hansen-Schirra 2021, 21).

With NMTs becoming more sophisticated and more accessible, how to evaluate MT output becomes an issue. Koehn (2020, 41–45) examines the issue from four different perspectives: task-based evaluation, real-world tasks, content understanding, and translator productivity. A phrase can be regarded as universally true is: "[q]uality is what the customer says it is" (Koehn 2020, 42), from which we can perceive the issue. The value of MT depends on the user. For some users, it is enough that the MT output conveys the intended meaning despite the language not being fluent, while for some purposes, fluency is also needed.

2.1.2 Post-editing

Post-editing has been cementing itself as a norm for almost all translation. Post-editing is defined as "the correction of raw machine-translated output by a human translator according to specific guidelines and quality criteria" (O'Brien in Nitzke and Hansen-Schirra 2021, 8). This means that a human translator edits the MT engine's output to make it more fluent and to correct possible grammatical, syntactical, and semantic issues. Also, many organizations that use these kinds of services have their own guidelines on wording and term usage, which almost always require human editing for them to be correct in the final product. PoE can be loosely thought exist in a domain somewhere between translation and proofreading. That being said, PoE seems to require a different set of skills than both fully manual translation and proofreading.

According to Nitzke and Hansen-Schirra translators began to use computers to aid with translation in the 1990s (Nitzke-Hansen-Schirra 2021, 9). The tools that the translators mainly used were some combination of "translation memory systems, terminology management systems and project management systems" (ibid.). TRADOS (translation and documentation software), which was established in the 1990s was the first TM system that was intended for commercial use (Nitzke and Hansen-Schirra 2021, 19). After the release of the first commercial NMTs, post-editing has more widely been conducted with the help of both an MT system and a TM. While NMT's improved fluency has been helpful for post-editors, Nitzke

and Hansen-Schirra suggested the following paradox, which makes the post-editors work more difficult as the MTs quality improves:

> the better the NMT translations are, the more difficult the error spotting is since the NMT output appears to be more fluent and less error-prone. This makes, on the one hand, the PE process even more demanding and leads to more cognitive effort for the post-editor. On the other hand, due to the absence of "real" errors, the post-editors tend to correct more style errors, which in turn leads to over-editing. (Nitzke and Hansen-Schirra 2021, 27).

2.1.3 The benefits of pre-editing

Pre-editing could be considered to be the most obvious solution for eliminating the concerns mentioned in the section above. In their 2020 book, Angelone, Ehrensberger-Dow, and Massey (2020, 334) define pre-editing as follows: "Pre-editing involves the use of a set of terminological and stylistic guidelines or rules to prepare the original text before translation automation to improve the raw output quality." They explain that pre-editors remove typographical errors, rewrite the text if the content is inaccurate, compact the text to form shorter sentences, rewrite the text to follow specified grammatical structures, clean the text semantically so that frequent terms are used, and clean out certain untranslatable content. Corpas Pastor and Durán-Muñoz (2018, 179) point out that there are numerous examples of pre-editing having been positively impactful in rule-based settings, statistical settings, and example-based settings. There has not been much study of how pre-editing impacts NMT settings, but with all that's been said of the NMT, there is a market for more research. In their 2017 paper, Koehn and Knowles studied how the output of an NMT compares to the output of a traditional SMT. They find that: "NMT systems have lower quality out of domain, to the point that they completely sacrifice adequacy for the sake of fluency" (Koehn, and Knowles 2017, 1) and that "NMT systems have lower translation quality on very long sentences but do comparably better up to a sentence length of about 60 words." (ibid.). Drawing from Koehn and Knowles' conclusions, we can establish that pre-editing could be very useful in NMT contexts. I can recall that, while only working for under a year in a company that provides language services, post-editing has been strongly compared to revision and their actual differences have been questioned by language professionals. If the pre-editing practices could be perfected, any kind of revision and proofreading would not be needed, because the MT

output would be as perfect as natural language can be. There is, of course, always a need for some kind of revision or proofreading for checking that nothing has gone wrong during the translation process, but the revisers workload would be significantly lower as there would be little to no errors in the text. Although, we again run into the problem that Nitzke and Hansen-Schirra (2021) proposed with PoE: Error spotting becomes harder when there are fewer errors and the revisers cognitive load may be increased as a result (Nitzke and Hansen-Schirra 2021, 27).

The most efficient way of pre-editing is through a controlled language (CL). Controlled language is defined by Huijsen as "an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style" (Huijsen in Marzouk 2021, 2). The construction and mining equipment manufacturer Caterpillar produced the now well-known Caterpillar Fundamental English (CFE) in 1972 (Mitamura and Nyberg 1998, 1–2). It was one of the earliest CLs that was developed for a specific use, that is, machine translating technical texts for, for example, the user manuals of mining machinery (ibid.). An improved version of the CFE, the Caterpillar Technical English (CTE), was later introduced (ibid.). These guides were used by the authors of the manuals and other similar texts, so that the texts could be translated into all the necessary languages without the need for human translation (ibid.). Caterpillar later discontinued the use of the CFE and the CTE because their product portfolio had grown so large and complex that the pre-editing guides did not provide sufficient guidance for such a wide array of technical texts (ibid.).

In their 2021 study, Miyata and Fujita studied what kind of results pre-editing could yield with a black-box NMT system. Here, black box means a system that can only be observed by its inputs and outputs with the inner workings of the system not being known. Fujita and Miyata's approach to this was to, first, input SL text to the MT engine and evaluate the engine's output on a 5-point scale of correctness. After the first step, the human controller executing the test would evaluate the correspondence of the source and the target and perform minimal edits to the SL text to if needed and, finally, select a version of the output that most accurately represented the SL text. The language pairs that were chosen to study were Japanese to English, Japanese to Chinese, and Japanese to Korean, and the MT engines that were chosen were Google Translate and TexTra (Miyata and Fujita 2021, 3-4). They found that pre-editing texts for an NMT can yield good results. However, they also found that contrary to the processes used when pre-editing for the earlier MT systems (RBMT and SMT) the results that the NMT provides do not benefit from shortening sentences and making them

more concise. Instead, the NMT's results benefit from the SL sentences containing more specific information overall and they do not seem to suffer from the sentences being longer (Miyata and Fujita 2021, 9). In addition, Miyata and Fujita conclude that NMT's are not, at least yet, consistent enough and they respond to very minor modifications in the source text by altering the output substantially (Miyata and Fujita 2021, 9). This lack of consistency suggests that they are not as well suited for translating controlled languages than earlier MT systems.

Ive et al. proposed an alternative approach to pre editing in their 2018 study. They studied if a human translator could benefit more from pre-translation than from post-editing. Their approach was three-pronged with the steps being the following:

> (1) automatic detection of fragments of the source text that could be problematic for the MT system; (2) resolution of these difficulties by a human expert, who provides the system with the expected translation of these segments; and (3) exploitation of the information by the by the MT system. (Ive et al. 2018, 2)

The difference between Ive et al.'s approach to the more traditional approach to PrE is that in their study a human translator would provide the MT engine with the appropriate translation rather than editing the source text in the SL and working with a specific set of rules according to which they modify the original text. It could be more appropriate to call this kind of operation pre-translating rather than pre-editing. They argue that, using this approach, the human translator has more control over the MT's function.

2.1.4 In-domain versus out-of-domain translation

One, typically thought of as a sociolinguistic factor called domain is also an important factor to consider in MT. In her 2002 book, Diana Boxer explains that "a domain refers to a sphere of life in which verbal and non-verbal interactions occur." (Boxer 2002, 4). The domain can, then, be considered as the language that is commonly used in a specific context and follows the lexical, grammatical, and syntactical norms of the said context. The scope of the domain can be modified to fit the context that is relevant to the study, e.g. technical texts versus the technical texts relating to industrial mining equipment. Of the two examples, the first one features a broader set of norms and the latter one features a more specified, narrower set of

norms. NMT engines are rarely fluent in-domain because of their expansive training data that typically consists of texts from multiple domains.

2.2 Differences in case between Finnish and English

As already introduced, Finnish uses a total of 15 case-inflections. The Finnish case inflections can be grouped into three different categories: grammatical, local, and marginal cases, with the nominative case not being grouped to any of these since it is regarded as a base form (Stephany and Voeikova 2009, 49–50). The local cases can also be divided into three subcategories: general, external, and internal local cases according to what their primary usage is (Stephany and Voeikova 2009, 51). Presented below are all of the 15 Finnish case-inflections including the nominative, their counterparts in English and their usage. Table 5 is partly based in my own knowledge of the Finnish language as a native speaker. The inflection suffixes and their English language counterparts are as they are presented by Huumo (Jaakola, and Onikki-Rantajääskö 2023, 18) and the usage descriptions are translated quotes from VISK (2008).

Table 1 The Finnish case-inflections, their respective suffixes with allomorphs, their English counterparts, and their most typical semantic usage.

Name of the case-inflection	Suffix	English counterpart	Usage
Nominative	-t (plural)	-	“Base form of a noun. Also used for various naming purposes such as in headings, captions, signs, tables, and lists” (VISK § 1231; my translation)
Accusative	-n,	-	“The accusative ending in ‘t’ has two constricted functions: it is the case of a total object and the case of the possessed in a possessive phrase in the inflection of seven words: pronouns and the interrogative pronoun ‘kuka’” (VISK § 1233; my translation)
Genitive	-n	‘of, ‘s’	“The genitive serves as a possessive case when used with a predicative and occasionally with the pre-modifier of a noun” (VISK § 1232; my translation)
Partitive	-a	-	“The partitive is typically used to mark the meaning of non-restriction. Non-restriction can mean indefinite quantity, group, or matter, or that

			the situation is presented as not having an end point." (VISK § 1234; my translation)
Essive	-nA	'as'	"The inessive, elative, illative, adessive, ablative, allative and the abstract locative cases essive and translative are locative cases." (VISK § 1235; my translation). "The inessive and the adessive imply that something is in a place, in a state, or in somebody's possession. The elative and the ablative imply that someone or something is moving away from a place, a state, or somebody's possession. The illative and the allative imply that someone or something is moving towards or into a place, a state, or somebody's possession. Of the abstract locative cases, the essive is grouped with the inessive and the adessive, and the translative is grouped with the elative and the ablative." (ibid.; my translation)
Translative	-ksi	'for, as'	
Inessive	-ssA	'in'	
Elative	-stA	'from, out of'	
Illative	-An, -On,-in, -en	'into'	
Adessive	-llA	'on, at, in the vicinity'	
Ablative	-ltA	'from on, from the vicinity'	
Allative	-lle	'onto, to the vicinity'	
Abessive	-ttA	'without'	"A phrase with the abessive indicates without what something is happening" (VISK § 1262; my translation)
Comitative	-ine	'with'	"The comitative indicates, firstly, a part of an entity that a subject or an object points to when the relation is descriptive; secondly, a comitative phrase indicates a person that is in someone's company or a social relationship meaning 'with whom or what'" (VISK § 1264)
Instructive	-n (plural)	'by'	"A phrase that contains with an instructive is primarily an adverbial phrase that indicates a manner or means" (VISK § 1263)

As can be seen from the above table, Finnish case-inflections are not the most straightforward features of the language to translate to English or vice versa, as many of the case-inflections are used in different ways in different contexts. In some cases, there can be free exchange in the use of case-inflections. This is most prominently the case between the inessive and the adessive, which represent two different locative case-groups in Finnish: internal and external. In addition to this, some case-inflections have special use cases that do not correspond to any of the ones mentioned in table 5. The fact that some case-inflections can be freely exchanged and that special use cases are prevalent does pose a problem for machine translation. The case that the MT chooses to use must be accompanied with a specific context and altering the context may result in the use of the wrong case-inflection. Although, now that we understand the basic principle in which the NMT operates, we can assume that it should have little to no problem in choosing the correct case inflections as the features that the NMT has encountered previously should weight all the features in a way that, when the output values are as close to

1 as possible, the translation should represented the most prevalent situation that the case-inflection has been encountered in by the engine.

2.2.1 Translating between languages with disparities in morphological richness

NMTs require large amounts of training data to translate accurately and a main problem in translating between synthetic, or morphologically rich, and analytic languages is that most synthetic languages are not among the most used languages in the world and there may not be sufficient amount of training data available to train an NMT. Although, according to Eberhard, Simons, and Fenning (2024) there are 7 164 languages spoken around the world today.

Tanwar and Majumder (2020) Explain that: “[n]eural networks have difficulty in modelling morphologically rich languages as compared to poor ones due to larger vocabulary, data sparsity, and added complexity (Tanwar and Majumder 2020, 4). They define morphologically rich languages as exhibiting a high degree of inflections, compounding or derivations” (ibid.). They add that: “In these languages, words are often modified to depict tense, mood, aspect, person, gender, number, modality, valency and so on” (ibid.). They mention Hebrew, Turkish and Dravidian languages as examples of MRLs (ibid.).

Translating between two MRLs is not an easy task. The affixes in MRLs do not necessarily behave in the same way, semantically, syntactically, or both. For example, one such language may use morphemes mainly as suffixes and another mainly as prefixes. MRLs can also often stack affixes behind each other, and they have to be used in a specific order for the word (or phrase as it could be considered) cannot be comprehended. Example of such a phenomenon is presented below.

Kirjasto	i	ssa	mme	kin
Library	plural	in	our	also

This example would, in English, be read as “In our libraries also” or “Also in our libraries”. Although, the Finnish word order is not very strict, the suffix order is very strict and has to follow the order: plurality identifier, locality identifier, genitive, “-kin/-kaan” conjunction (meaning “also” or “neither”), question identifier. This complicates the MT systems job even further because it has to understand which category each suffix belongs to and assign them to the correct place following the presented order.

One argument of the MT systems’ difficulties in translating from or to MRLs is that, according to Tanwar and Majumder (2020, 4), training data from these languages is scarce and not enough for the MT systems to learn. A more significant factor could be thought to be the multi-faceted systems that many MRLs are based on.

2.2.2 The prevalence of case-inflections

Different cases appear in different frequencies in the Finnish language. Institute for the languages of Finland has gathered data from a selection of natural language repositories to form a grammatical archive. One of the many applications of such wide array of data is to calculate the frequency in which each case-inflection appears in Finnish. The observations are presented in the table below.

Table 2 The findings of VISK § 1227 from the grammatical archive corpus and the Parole corpus.

	Grammatical archive		Parole corpus	
	%	n	%	n
NOM	30,5	39939	37,4	59251
GEN	22,9	29912	21,7	34393
PAR	13,6	17734	13,7	21720
ACC	0,1	108	0,1	107
ESS	2,6	3434	2,1	3393
TRA	2,2	2895	1,6	2546
INE	6,8	8864	5,8	9200
ELA	4,6	6072	4,0	6316
ILL	6,7	8792	6,1	9606
ADE	4,2	5553	4,0	6355
ABL	1,1	1457	1,0	1490

ALL	2,4	3123	2,2	3435
ABE	0,2	308	0,2	266
KOM	0,1	120	0,1	103
INS	2,0	2610	0,3	474
TOTAL		130921		158655

Table 6 shows the results that were gathered from the grammatical archive and Parole corpus, which is a mechanically produced sample of a corpus based on articles found in Helsingin Sanomat (Finland's largest non-national news media) and Suomen kuvalehti (Finnish magazine), and an unnamed non-fictional book (VISK § 1227). From the table's contents we can derive assumptions on how much different cases are used in the Finnish language. From the data we can see that some case-inflection suffixes are used exponentially more than others. Of the four basic cases, nominative, genitive, partitive occur much more than any other with nominative being the most frequently used. Next most used are the locative cases: Essive, Translative, Inessive, Elative, Illative, Adessive, Ablative, Abessive, Allative. There are still quite large discrepancies in the frequencies of usage between the locative cases, for example: the ablative (1.1 % and 1.0 %) and the illative (6.7 % and 6.1 %). These discrepancies may be the result of two things: The more frequently used case-inflection refers to phenomena that are more frequently discussed or the more frequently used has a special use case, which allows it to be more versatile and, as such, more appropriate to use in more situations. Finally, the third category of cases, which KOTUS refers to as Cases of incomplete use (Finnish: Vajaakäyttöiset sijat): abessive, comitative and instructive. These are referred to as cases of incomplete use because their use entails lexical and morphological constraints (VISK §1227). The incidence of these cases is low (below 1%) except in the grammatical archive's corpus, where the instructive comprises 2 % of the data. The unexpectedly high incidence is explained as being the result of some adverbs and particles having been coded so that they include the instructive case (ibid.). This can be regarded as a special use case and, as such, will not be taken into account in this study.

2.3 Quality metrics

The evaluation scale that is used to analyze the results in this thesis has been created by modifying Multidimensional quality metrics (MQM), which is a set of metrics designed to measure translation quality. However, MQM is an extensive set of metrics that are designed to analytically examine different kinds of problems within a translation. The main superset of components within MQM are the following: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup (MQM, n.d.). Considering the quality of the study, most of the components can, here, be disregarded, and the ones needed are accuracy and linguistic conventions. MQM also, in its usual format, features a passing score, which has not been implemented in the evaluation scale used in this study, since giving passing scores to such short texts that feature issues would not be desirable. Also, because the sample size is too small to give comprehensive results, the aim of this research is not to determine how the MT engines rank in comparison with each other but to analyze the common types of mistakes that they make and to hypothesize the possible reasons for the mistakes.

Other quality metric system that is meaningful in researching MT performance is the bilingual evaluation understudy (BLEU) scoring system. BLEU is a scoring system that has been specifically created for evaluating the quality of machine translation and as such it is not a good tool for measuring translation quality. BLEU measurement uses a set of human translation as a reference and cross-references the MT output segment to the human translated segment or segments. The BLEU measurement then assesses how many words in the MT output match the human-translated reference material with subsequent correct matches resulting in a higher score than individual correct words. The primary reason the BLEU scoring system is not great at evaluating translation is that it does not consider, for example, synonyms or paraphrases. (Vashee, 2019). BLEU system is currently widely used in computer-assisted translation, or CAT, tools, for example Trados Studio, which cross-references the MT translated segments to adjacent matches from the translation memory that is used.

3 Materials and Methods

In this thesis, I will use samples of text that are structured in a way that their Finnish translation will most preferably feature any one or more of the case-inflections that are the focus of this thesis. The samples are gathered from varying sources to account for different styles of text.

3.1 The text samples

For this thesis, the texts are gathered from BBC news, Lewis Caroll's fictional book Alice's Adventures in Wonderland, and Manjit Dosanjh's scientific book From Particle Physics to Medical Applications The specific samples were chosen for this study by scanning texts for possible presence of structures, whose Finnish translation could include the desired case-inflections, with the help of DeepL machine translation engine. Each sample contains one structure that, when translated to Finnish, should contain a case-inflection if the translation is correct in other ways.

Table 3 Text samples from the News-texts section

Partitive	"It cements her position as the one of the greatest songwriters of her era" (Savage, 2024)
Essive	"As a child, she relished simple rituals of the Lunar New Year" (Guo, 2024)
Translative	"I was trained as a conservationist and you do not interfere, you don't move things to new locations." (Iovenko, 2024)
Inessive	"Lana Del Rey, calling her 'a legend in her prime'" (Savage, 2024)
Elative	"Its aim is to discover new particles that would revolutionise physics and lead to a more complete understanding of how the Universe works." (Ghosh, and Stephens, 2024)
Illative	"Its aim is to discover new particles that would revolutionise physics and lead to a more complete understanding of how the Universe works." (Ghosh, and Stephens, 2024)
Adessive	"At the time of release, it was not her best-received album." (Savage, 2024)
Ablative	"The final tier included residents who were displaced from the island prior to Hurricane Isaac." (Sherriff, 2024)
Allative	"The existence of a building block that gives all other particles in the Universe their form was predicted in 1964 by the British physicist, Peter Higgs" (Ghosh, and Stephens, 2024)

Table 4 Text samples from the fictional- texts section (Carroll, [1865] 2000).

Partitive	But do cats eat bats, I wonder?
Essive	burning with curiosity, she ran across the field after it, and was just in time to see it pop down a large rabbit-hole under the hedge.
Translative	it was labeled "ORANGE MARMALADE," but to her great disappointment it was empty.
Inessive	So she was considering, in her own mind
Elative	the Rabbit actually took a watch out of its waistcoat-pocket.
Illative	she found herself falling down what seemed to be a very deep well.
Adessive	yes, that's about the right distance—but then I wonder what Latitude or Longitude I've got to?
Ablative	There seemed to be no use in waiting by the little door, so she went back to the table, half hoping she might find another key on it, or at any rate a book of rules for shutting people up like telescopes: this time she found a little bottle on it.
Allative	she was walking hand in hand with Dinah, and was saying to her, very earnestly.

Table 5 Text samples from the scientific-texts section (Dosanjh, 2017)

Partitive	On 4 July 2012, both the ATLAS and CMS collaborations announced that they had observed a new particle consistent with the Higgs boson.
Essive	This can serve as a model for emerging multidisciplinary ventures in medical applications.
Translative	The work leading to the discovery—what The Economist lauded as ‘science’s great leap forward’ [2]—represented the culmination of decades of effort.
Inessive	This was almost 50 years after the particle had first been predicted in theoretical calculations by Peter Higgs, Robert Brout and François Englert.
Elative	This mode of working has become second nature for particle physicists, who have learned to work collectively towards a common goal and who rely on consensus to take decisions.
Illative	It is clear that physics, and in particular particle physics, has made a major contribution to the development of instrumentation for biomedical research, diagnosis and therapy.
Adessive	CERN initiated a study to review the available technologies and determine what further developments would be needed to meet the requirements of this emerging treatment modality.
Ablative	These two independent detectors exploit different technologies, which is crucial for crosschecking and confirming any new discoveries.
Allative	Physicists, engineers and computer scientists can share their knowledge and technologies, providing the medical community with first- hand information on the latest technical progress.

The dataset does not include all 15 of the Finnish case-inflections. This can be explained by their extremely high or low frequency. The two most basic cases, Nominative and genitive, are not included in the dataset because of their extremely high frequency in typical language use and their existence in both English and Finnish. Because of this, we can assume that the MT engines translate them with almost 100 % accuracy. Accusative, abessive, comitative, and instructive are not included either. This is due to their extremely low frequency, which can be assumed to cause the MT engines to translate them with almost 0 % accuracy. The special use cases are disregarded in the data gathering phase. This is done due to the special use cases being special cases and, at this moment, it is clearer and simpler to use the most basic use cases as data so that the data will be reproducible.

The samples were gathered from articles by Ghosh and Stevens (2024), Savage (2024), Guo (2024), and Iovenko (2024) on the BBC News website, Lewis Carroll's ([1865] 2000) classic book Alice's adventures in wonderland, and Manjit Dosanjh's (2017) book From particle physics to medical applications, by feeding chunks of text into DeepL translator and picking out phrases that were translated to have the desired case-inflection. The length of the context that was included was determined by the assessed need for context. For example, in the fictional-texts section, in the ablative case, the word that the case-suffix modifies is the auxiliary word "it" that refers to the word table that appears around 30 words before. I assumed that the MT engines would produce the best and most accurate results when the samples, that they were given, were as clean as possible and no, so-called 'noise' was present. The samples were gathered from news media, a fictional work, and a scientific textbook, so that the most prominent registers would be included, and we would be able to examine if there are any differences in the MT engines performance between these domains. There are, of course, more domains, in which MT engines performance could be researched. I chose to focus on news media, fictious works, and scientific works, because scientific and fictious styles are almost completely opposite with news media style featuring qualities from both. This way the results should not be too vague and the examination of the reasons for the MT engines performance does not become ostentatious.

3.2 The MT engines

The main focus of this thesis is to test the capability of Google's, Amazon's, Microsoft's, and DeepL's MT engines in translating English structures to Finnish and choosing the correct case-inflection for the purpose. The reason for choosing these 4 MT engines is that they are the most prominent in today's translation landscape, both in the industry and within the everyday use of the general public. All four MT providers offer free of charge translations for a limited amount of text with some caveats, such as having to be subscribed to a service. The MT engines were accessed via Phrase TMS, an online translation workflow software, because Phrase TMS allows the user to pick any one of the aforementioned MT engines to a specific project. This made the translation part of the study much less complex as all of the work could be done on one platform. The text samples that were entered into Phrase TMS were the ones that are shown in tables 1, 2, and 3 with no context added. The decision to not add context to any of the samples was made to ensure that all of the samples had only the necessary context and nothing more. Also, I assumed that adding context could make the results more convoluted. Some samples are from the same sentence and have been split into the samples that are presented in tables 1–3.

3.3 The evaluation scale

After the translation, the samples were collected, and they were analyzed with the help of a seven-levelled evaluation scale with which the translated samples could be assigned one of the numerical grades presented in table 4.

Table 6 The evaluation scale

Numerical grade	The issue that numerical grade relates to
0	The target text does not feature any case-inflection and also does not correspond to the subject matter of the source text.
1	The target text features a case-inflection that does not correspond to the meaning of the structure in the source text.
2	The text features the correct case-inflection, but in a wrong position, thus changing the meaning of the text.
3	The target text features a case-inflection, which is not a preferred one, but only slightly changes the meaning of the text.

4	The target text does not feature a case-inflection. However, the target text matches the source text in subject matter.
5	The target text features a case-inflection that is not the preferred one, but the change does not significantly alter the meaning of the text.
6	the target text features the desired case-inflection and otherwise matches the source text in subject matter.

This evaluation scale has been constructed as a response to an earlier test performed with an evaluation scale that was similar in form but only featured three grades. The test was performed by me with the purpose of examining the validity of the research topic and methods. The three grades that were chosen for this test were 0 = the case-inflection was not translated, 1 = the case-inflection was translated incorrectly, and 2 = the case-inflection was translated correctly. The limited grading system lacked sufficient nuance and resulted in the results being too narrow in scope, which is why the evaluation scale was updated to the seven-levelled form. The evaluation scale is loosely modelled after the MQM scale. However, the evaluation was made to be criteria-based because the primary focus of the evaluation is if the translations include a case-inflection and the secondary focus being on how correct the choice of case-inflection is in context. This two-fold evaluation can most accurately be conducted with a criteria-based evaluation scale where both foci are considered in the criteria.

The grades have been designed to serve more wide-scoped purposes than visible on first glance. The highest grade is quite self-explanatory as it expresses an absolute success. The lowest grade, however, as well as expressing the pseudo-numerical value of absolute failure, may also indicate that the presence of a (most likely rare) case-inflection in the target language has confused the MT so much that it has not been able to produce a sensible translation at all. The grades are also not designed to be in a hierarchical order of betterness. For example, a translation of a sample that has been assigned the score of one, because it does not feature any case-inflection can, in a real-world-scenario, be preferred over a target sample that has been assigned a score of three. This is because the evaluation scale has been specifically constructed to give accurate results in a very narrow scope and the scale's accuracy starts to suffer when trying to apply it to a bordering subject.

3.4 Qualitative examination of the results

After the initial test had been performed, and the results had been evaluated according to the evaluation scale, a qualitative examination of the results was conducted. As with all quantitative research methods, the evaluation scale featured in this research is highly limited in assessing the true nature of issues when the subject is something with a highly nuanced set of variables, such as language. The evaluation scale, being designed to feature a set of seven different values, cannot present all of the information needed to accurately assess every factor contributing to the correctness or incorrectness of the translation.

The values provided by the evaluation scale were more thoroughly examined by contrasting with the frequency of occurrence of each Finnish case-inflection provided by the web version of the book Iso Suomen Kielioppi (VISK § 1227), after which the possibilities of pre-editing being the solution to the problem were examined. The difficulty for discussing possible MT features that could be beneficial in the translation of case-inflections is that the NMTs that are featured in this research are black box MT systems, at least for myself. Black box means that nothing other than the inputs and outputs can be observed and whatever happens inside the system is not visible. The models that DeepL's, Google's, Microsoft's and Amazon's NMTs use are proprietary and trade secrets, which cannot be acquired for the purposes of publicly available research. However, the fact that the NMT infrastructures concerned here are mostly a mystery for anyone except people who work with them, brings the possibility of creating ideas that, while maybe not useful with the current infrastructures, could be useful to future infrastructures that are created in the field of language technology.

4 Results

4.1 The presentation of the results of the test of the Machine translation engines' capability of translating case-inflections correctly

The results of the evaluation of the translated samples are presented in the following tables 7–9. The numerical values are assigned to each translation according to the ruleset presented in section 3, 0 being the lowest and 6 being the highest possible grade. The grades are designed to only indicate whether the case-inflection is correctly translated in the corresponding translation and do not express the overall quality of the translation.

Table 7 The results of the evaluation from the news-texts portion.

	DeepL Translator	Google Translate	Microsoft Translate	Amazon Translator
Partitive	6	6	6	6
Essive	6	6	6	6
Translative	6	6	6	6
Inessive	6	3	3	3
Elative	6	6	6	6
Illative	6	6	6	6
Adessive	6	6	6	6
Ablative	6	6	6	6
Allative	6	6	6	6
Total score (average)	6	5.66..	5.66..	5.66..

Table 8 The results of the evaluation from the fictional-texts portion.

	DeepL Translator	Google Translate	Microsoft Translate	Amazon Translator
Partitive	6	6	5	6
Essive	6	1	6	3
Translative	6	5	6	5
Inessive	6	6	6	6
Elative	6	6	6	6
Illative	6	1	3	4
Adessive	6	2	2	2

Ablative	6	0	3	3
Allative	6	6	6	6
Total score (average)	6	3.66..	4.77..	4.55..

Table 9 The results of the evaluation from the scientific-texts portion.

	DeepL Translator	Google Translate	Microsoft Translate	Amazon Translator
Partitive	6	5	5	5
Essive	6	6	6	6
Translative	6	6	6	6
Inessive	6	6	6	6
Elative	4	4	6	6
Illative	6	6	5	6
Adessive	6	5	5	5
Ablative	6	4	5	5
Allative	6	6	6	6
Total score (Average)	5.77..	5.33..	5.55..	5.66..

When examining the output of the MTs, I found multiple surprising issues. For example, DeepL was, in some cases, not able to reproduce the same target text as it had produced when gathering for the data, that were used in this test, that featured the case-inflections. The changes in output were subtle and not at all remarkable even if the text was to be published to an audience. Nonetheless, the surprising aspect is that there seemed to be nothing in the context, that was left out, which could be considered influential for these changes. DeepL presented no issues with the fictional text, which was overall the most difficult for these MT engines. This may be due to the more idiomatic and highly specialized nature of the text. DeepL may simply have not had any better or even equally good options to choose from, so it had to translate the text in the same way both times.

4.2 The inspection of the results

All of the MT engines performed extremely well when translating the news texts (average grade 6 for DeepL and 5.66.. for the other three), which is not very surprising since news texts are prevalent and easily accessible which means that there is a great probability that texts in the specific domain are highly represented in the MTs' training data. The MT engines also performed well when translating scientific texts (average grades between 5.33.. and 5.77..) . The fact that the engines struggled when translating the fictional texts examples is not surprising and may be the result of two main factors that are crucial in determining how well MT engines perform: Sentence length and the idiomaticity of the used language. Although, by observing the samples we can observe that the average sentence length of the scientific texts is considerably higher (approx. 22.66 words) than that of the fictional texts (approx. 17.77 words) and the news texts (approx. 16.66 words). Here, the average length is not an appropriate predictor of accuracy, since, when considering the median length of sentence between the sample groups (news = 15, fictional = 13, and scientific = 23) the difference in median sentence lengths between them is not considerable enough. We must, then, examine the variance in sentence length. The news texts had a variance of 16 words, with the shortest sample being 10 words long and the longest being 26 words long. The fictional texts had a variance of 44 words, with the shortest sample being 7 words long and the longest being 51 words long. The scientific texts had a variance of 16, the same as the news texts, with the shortest sample being 13 words long and the longest being 29 words long. We must then consider that, according to Koehn and Knowles' (2017, 1) findings, NMT systems perform poorly when sentence length exceeds 60 words. None of the samples in this study exceed the 60-word sentence length threshold that Koehn and Knowles suggest being a predictor of poor performance for the NMT. Then we must consider the opposite: have the MT engines had enough context for them to perform properly. Shorter sentence length may hinder, because the MT engine does not have enough material to work with. Although, when gathering the sample data and entering the samples through the MT engines it seemed that a necessary amount of context was provided to them. Still, it may have improved the scores if more context was provided to the MT engines. This problem also involves the complexity of the text. For instance, fictional texts often include, for example, idiomatic expressions and metaphors, which may not follow the conventional style of language, hence producing problems for the MT.

By this evaluation method, Google translate was the least accurate and DeepL the most accurate MT engine. Although, we have to consider that the setting is biased to DeepL's benefit, as the examples were gathered using DeepL. In a more equal setting, the results may be more balanced or there may be more variance between the engines' performance. Different results may be obtained by changing the way the data is gathered, by expanding the scope of the evaluation scale, or by altering both. Google was not clearly the worst with its average score being less than one point worse than all of its competitors in the news-texts and scientific-texts sections, and approximately 1,11 and 0,89 points worse than its key competitors, Microsoft translate and Amazon translator, respectively in the fictional-texts section. We must also take into account that the evaluation scale, featuring just seven values that are assigned to the results, may not be extensive enough to accurately portray the consequences of these inaccuracies in real-life language use.

4.3 Preliminary observation of the issues

None of the MT engines presented major problems in sentence or word structure when using the case inflections. In some cases, the incorrect usage of a case-inflection seems to affect fluency in a significant way, such as in example 1.

Example 1. *(Original)*

she found herself falling down what seemed to be a very deep well.

Hän huomasi putoavansa alas, mikä näytti olevan hyvin syvä kaivo.

The causal connection between the incorrect case-inflection and disfluency can only be hypothesized since we do not have an inside look into the operation of any MT engine. We can assume that the problem lies somewhere other than the use of a case-inflection particle since they are such a small part in a complete sentence structure. Nevertheless, it is not impossible that such a small particle as a case-inflection caused the whole sentence to be incorrectly translated. If we observe the basic operation of an MT engine presented in section 3.1.1, we can roughly estimate the probability of this happening. A MT engine processes each particle as an input node. We do not know how big the specific particles are and how their relationships have been coded to influence the output. In his 2020 book, Philipp Koehn (2020, 126) says that an input node corresponds to a word. Words are the smallest grammatical units in, for example, the English language, but not in Finnish or other MRLs. This could mean that

the NMT engine may not account for a specific grouping of nodes when translating, which leads to problems, such as leaving out specific morphemes or replacing them with other ones that seem similar and share the probability of usage with the correct ones. The issue may be caused be low specificity. Much of the prior data, such as that presented by Koehn and Knowles (2017), shows that NMTs accuracy is poor when translating longer sentences. Koehn suggests that this failure is an infrastructural problem that happens due to the NMT not being able to remember the first words in a sentence when it executes the decoding phase (Koehn 2020, 126). When examining the result of the ablative case in the fictional-texts-part of the study (see appendix 2.), we can assume that this exact infrastructural problem is the reason for the MT engines' poor performance. The sentence is very long, and the ablative case is supposed to be presented in the last word of the sentence, which refers to the word "table" that has been presented in an early part of the sentence. If this assumption is correct, the MT engines cannot be said to be poor at translating the ablative case, but simply poor at translating the later portions of longer sentences. The same cannot be said about the illative case in the same section, which was also translated poorly by the three MT engines that were first timers in translating the sentence. The sentence, where the illative case is present, is short but the MT engines still had trouble with translating to the illative case (see appendices 1, 2, and 3). Illative case is also not infrequent in Finnish text as shown by Hakulinen et al. (VISK § 1227). This means that there is no apparent reason why the MT engines were not able to accurately translate the illative case. In favor of the illative, Google translate used the ablative, Microsoft translate used the elative, and Amazon translator dropped the case-inflection altogether. None of these choices resulted in the sentence being semantically correct. The sentences that Google and Microsoft produced were still syntactically and grammatically correct, but Amazon's sentence did not fulfill any requirements of a good translation. Although, the sentence that Amazon produced was closer to the correct meaning and, if a translator were to post-edit the sentences, Amazon's translation would require less editing than the others. The issue that the MT engines faced in this instance can be explained with the fact that the sentence is from a fictional work and the language is freer than it would be in a formal text. The sentence lacked the defining preposition "to", which is needed for the MT to assess what the case is. That said, DeepL had no issue with this omission. The omission can be disregarded if the sentence is examined at a higher level. This means we can assume that DeepL is more sophisticated in the regard.

When examining the Scientifical-texts-portion of the study, we can see that in the case of the partitive no MT engine other than DeepL chose to use the partitive case-inflection (see appendix 3). The three MT engines chose to use the genitive. This seems to be the result of the engines translating the sentence in a more direct fashion. Google's, Microsoft's, and Amazon's translations follow the structure of the source text more accurately than DeepL's translation. DeepL's translation seems more fluent to a human reader but, at the same time, does not convey all of the information presented in the source text. Nonetheless, this does not mean that DeepL was wrong in choosing to use the partitive case instead of the genitive. The sentence can be translated using either case and upholding the semantic connection to the source text and conveying all the information conveyed by the source text.

The only issue for the MT engines in the news-texts-section was the inessive case. The sentence was highly idiomatic and DeepL was the only MT engine that was able to translate the idiomatic expression correctly. The other three engines chose to modify the meaning from "in her prime" to "at her best" (see appendix 1). However, this could be translated in such way by a human translator because the idioms do not exactly match semantically between English and Finnish. Overall, DeepL seemed to have a more author-like approach to translating this sentence, inserting embellishments despite them not being present in the source text. DeepL seems to have been made to edit the text itself without the need for a human language professional to check what the engine has produced, or so it may have been intended.

4.4 Does the frequency of specific case-inflections affect the MTs' accuracy?

The news-texts-section featured only one case with which the TM engines had trouble, the inessive (see appendix 1). The fact that the MT engines had trouble with the inessive is remarkable in that it suggests that the frequency of the case-inflections' usage has little to no connection to how well the MT engines can use them. Although, in this case it seems that the MT engines' choice to not use the inessive seems to be intended as simpler than how DeepL translated the sentence. The MT engines swapped the inessive to adessive so that they did not add the noun which the inessive suffix modified and had the adessive modify the adjective that is present in the original text. DeepL produced a more eloquent translation by adding the noun in the inessive case. This cannot be regarded as a complete failure on Google's, Microsoft's and Amazon's part because they simply translated the sentence that they were provided with great accuracy.

In the fictional-texts-section, we can observe more correlation between the cases' frequency of usage and the MT engines' accuracy in using them (see appendix 2). First such case is the essive, which was not translated as expected by Google and Amazon. Essive is not the most infrequent case-inflection examined in the study with its frequency being 2.6 % and 2.1 % in the Grammatical archive's corpus and the Parole corpus respectively. Nonetheless, it is still infrequent enough to be predicted to be troublesome for the MT engines. Google had a hard time with translating the essive sentence, seemingly trying to swap out the essive for an instructive modifying an adjective and also failed in that effort. This is very unexpected since the instructive is one of the cases that was dropped out from the study altogether because of its low frequency of use. That said, the phrase in which the essive was supposed to appear in was very idiomatic, which can be expected to result in a corresponding idiomatic translation. In addition, Google did not succeed in using the instructive, which further supports the hypothesis that the MT engines will not be able to accurately translate the low frequency cases at all. Google seemingly tried to translate "burning" to "palaen", which would be a correct translation but added an extra letter "t" to make the word "palaten", which means "returning". Amazon, on the other hand, missed the idiomaticity completely and altered the meaning of the phrase when translating. Amazon also favored the instructive over the essive and surprisingly succeeded in using in the correct form. Nevertheless, Amazon modified the phrase semantically so that the word "burning", which would correctly be translated as "palaen" (meaning: something is burning), was translated to "polttaen" (meaning: someone is burning something). This, quite interestingly, contradicts the earlier assumption and the supporting argument that the MT engines cannot translate low frequency cases accurately and, when examining the variables, the phenomenon becomes even more peculiar. The translated phrase was very idiomatic in the SL, but Amazon translate seemed to have not understood the idiomaticity and translated the phrase in the same way as any other phrase, yet it was able to use the instructive in a way that was correct in structure but not correct when observed in context. It seems as though nothing prompted Amazon to use the instructive except the idiomatic nature of the phrase, but Amazon disregarded the idiomaticity by using a semantically incorrect word and still deciding in favor of the instructive. The issue may be that the most problematic word appears in the initial position of the sentence and, if the MT engine is not familiar with such idiomatic phrase, it is not able to reverse process it so that all output nodes have a high enough probability to be accurate. Phrase TMS, the web-based translation tool with which the study was conducted, provided the phrase with a translation quality score of 77 %, which is quite high considering the quality of Amazon's mistake.

However, we must take the length of the sentence, which was 26 SL words, into account. The sentence may be otherwise extremely well translated and all of the nodes except the first one had a high probability, which raises the suggested accuracy.

The next, and the fourth most problematic case in the fictional-texts-section was the illative, with which the MT engines other than DeepL struggled immensely (see appendix 2). The sentence that they had to translate was “she found herself falling down what seemed to be a very deep well”. Google and Microsoft did not seem to understand that a well is something that you can fall into and translated the phrase so that the well was something from which the person falls. Interestingly, Google and Microsoft used different cases in their translations. Google used the ablative, which implies that the person fell down from somewhere nearby the well, for example, from on top of it, while Microsoft used the elative, which implies that the person fell from the inside of the well. Both the elative and the ablative occur less than the illative with the elative appearing 4.6 % and 4.0 % of the time, the ablative appearing 1.1 % and 1.0 % of the time, and the illative appearing 6.7 % and 6.1 % of the time in the Grammatical Archive’s corpus and the Parole corpus respectively. The difference in the frequency of occurrence between the elative and the illative is quite small 2.2 percentage points in both corpora, which is minimal and the fact that Microsoft chose to use the elative can be considered coincidental. The difference in the frequency of occurrence between the ablative and the illative, however, is 5.6 percentage points in the Grammatical archive and 5.1 in the Parole corpus, which is considerable and suggests that there has to be a reason for Google translate to have chosen to use the ablative instead of the illative, which would have been more appropriate. Amazon translate, while having decided not to use any case-inflection, was still semantically more accurate in its translation. Again, Amazon translate seemed to translate the phrase with a more straight-forward approach than the others, which can be considered as both a virtue and a vice, since it is semantically more accurate but, because the grammar and syntax are not as accurate as the others, it does not offer the same help as the others for a post-editor. One aspect that may have had an effect on the translation choices made by the engines is that the phrase does not feature a traditional preposition that would dictate what case-inflection should be used in the translation. In the phrase, the word “down” acts as the preposition that should dictate what case-inflection should be used. This may have been confusing to the MT engines and they chose the case-inflection that they deemed the most appropriate.

The adessive seemed to be one of the two most difficult case-inflections for the MT engines (see appendices 1, 2, and 3). Google, Microsoft, and Amazon all decided to use the adessive in the following word from where it was excepted and, additionally, Google and Amazon decided to use the illative in the word where the adessive was expected to appear in. While, in accordance with the evaluation scale, all three engines have been assigned a value of two from this part, the scores do not accurately reflect the performance of the engines. Microsoft outperformed Google and Amazon because it did not add the, in this instance, undesirable illative. The MT engines could be expected to use the illative instead of the ablative because of its higher frequency of occurrence but using them both in succession is peculiar. The decision seems to be in accordance with the argument that the NMT's accuracy is worse towards the end of sentences (Koehn 2020, 126). It has been suggested that this is true with longer sentences, but in this instance, the source sentence in not very long, 17 words long, which cannot really be considered as a long sentence. A more probable reason for this inaccuracy is the fact that the last phrase of the sentence features what Richard P. Laverdure (1983, 1) calls a hanging preposition. A hanging preposition occurs when a preposition occurs in the final position of a phrase. Here, the preposition "to" is left in the final position of the last phrase and the MT engines seem disregard its existence completely.

The ablative was, along with the adessive, the most problematic case-inflection in the fictional-texts-portion. Microsoft and Amazon found a tolerable substitute for the ablative in the elative, while Google decided to use the illative. Elative is an acceptable case to use in this instance since it implies movement that is directed outward in the same way as the ablative while the illative implies movement that is directed inward. The elative and the illative are both considerably more frequent in the corpora than the ablative with the elative being 3.5 and 3.0 percentage points more frequent and the illative being 5.6 and 5.1 percentage points more frequent than the ablative in the Grammatical archive's corpus and the Parole corpus respectively. It would be plausible to assume that the low frequency of occurrence of the ablative is the primary reason for the MT engines opting to use other, more frequently occurring case-inflections in favor of the ablative. However, as with the earlier examples we must take the length of the sentence into account. The source sentence is 52 words long, which is a considerable length for a sentence and Koehn and Knowles' (2017, 1) finding that NMTs accuracy drops significantly when translating longer sentences, is supported by this finding.

In the scientifical-texts-section of the study, all of the phrases were translated with acceptable accuracy with a few issues of a low or very low severity being presented. We can assume that the scientific text style features such strict grammatical and syntactical rules that it is easier for the MT engines to translate that text style. The problems of lower severity are ones where the MT engine has modified the syntax from the one that was expected and thus has needed a different case suffix or no case suffix at all. These cannot even be regarded as problems, because the syntax, grammar and semantics of the MT outputs are satisfactory.

4.5 Problems with morphological richness

None of the MT engines displayed serious difficulties with the morphological richness when translating from English to Finnish. Nonetheless, there were examples where the difference in morphological richness may have caused difficulties for the MT engines. These examples are mainly found in the fictional-texts-section as it was the most difficult for all of the MT engines.

First such case could be in the essive case. The phrase that was the main focus of this study was “burning with curiosity”, which was expected to be translated “uteliaisuudesta palaen”. We can observe that the SL phrase is made up of three words and the TL phrase is made up of two words. Despite this, both phrases consist of the same number of morphemes, four. We can conclude that the MT engine is expected to not only understand how the SL and TL morphemes correspond to each other but to also understand their relationship with the other morphemes within the phrase in each language. The basic relationships are illustrated in figure 1.

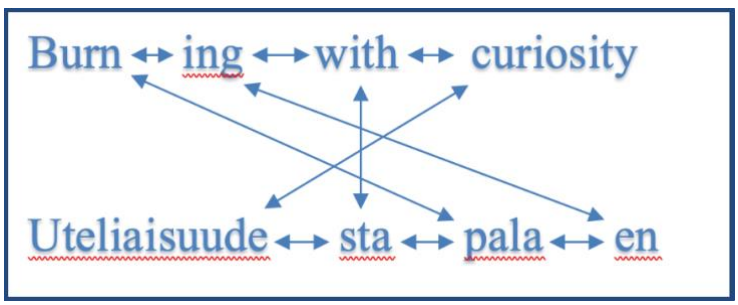

Figure 1. The semantic equivalence of the constituents of the phrase “burning with curiosity” and its predicted Finnish translation.

Figure 1 shows the most basic relationships between the morphemes, which can be illustrated in a two-dimensional figure. If the interchangeability between Finnish case-inflections was to be added to the model, the model would have to be three dimensional, with all the plausible case-inflections being on the Z-axis side by side perpendicular to the horizontal X-axis. All of the plausible case inflections in the TL would be connected to their SL predictor, as well as the words that are predicted to appear before or after them. In a typical situation, the translation of a phrase that follows the form "with something" would not be predicted to feature a case inflection at all, the preferred translation featuring the postposition "kanssa", which is the more frequently used counterpart to the comitative in Finnish. In this scenario, the phrase is idiomatic and the word "with" does not correspond to its natural use, which requires also requires adaptation of grammatical forms in the TL by the MT engine. Thus, the MT engine faces multiple issues with the translation of a phrase such as this. Neither of the TL case-inflections are ones that would typically be used in a situation such as this.

In the case of the illative, it seems as though the MT engines lack the context that they need. The phrase is not long but the SL morpheme that is the most important predictor to the case-inflection, that is needed in the end of the TL phrase, appears in the first half of the text. There is also a subordinate clause in the middle position of the sentence that is not needed, and not even included in the translation by Google translate, that might be a reason for the MT engines struggling with the translation. Koehn (2020, 126) explains that MT engines struggle with longer sentences because it cannot "remember" the start of the sentence when it reaches the end and cannot take it into account when translating. This should not happen in a phrase consisting of 13-word-long sentence, but it seems as though this is exactly the problem with this example.

In the case of the adessive, it seems that the MT engines also struggle because the language is not grammatically similar to their training data. Fictional texts often feature more relaxed grammatical rules than other text types and that leads to phrases being constructed differently in fictional texts than in other types of texts. The sample has a hanging preposition in the final position. Hanging prepositions occur very frequently in speech and more informal language use. The fact that the MT engines, other than DeepL, seem to not be familiar with the use of hanging prepositions suggests that their training data features little to no informal content, which may be by design. When translating formal texts, for which at least Amazon's and Microsoft's engines are probably designed for, it is not desirable for the MT engine's output

to include informal grammatical, lexical, or syntactical features such as the hanging preposition.

5 Discussion

5.1 Could pre-editing solve the accuracy problems?

Miyata and Fujita (2021, 8) found that structural pre-editing changed the MT engine's output significantly. To pre-edit a phrase so that it is easier for the MT engine to understand and translate, the pre-editor must modify the structure of the text so that the SL words that correspond to the case-suffix and the word that the case-suffix modifies are in specific places so that the MT engine can understand that they are meant to be in a modifier–modified relationship. I propose that the easiest way to achieve this could be to make sure that as few words as possible exist between the modifier and the modified. We can observe from the gathered data that the MT engines had significant trouble if the modifier and the modified were separated by more than one word within a phrase. Miyata and Fujita also found that "many of the editing types that include local modifications of functional words and orthographic notations … did not have major impacts on the MT results" (Miyata and Fujita 2021, 8). Local edits could be difficult at least for some cases where case-inflections lead to problems in translation output. Nevertheless, there is a possibility of implementing both structural and local modifications simultaneously being able to provide improved results. Although, this does not remove the problem of structural edits resulting in major alteration in the MT engine's output.

An aspect that must be considered when examining the use of PrE is the domain. If the NMT cannot determine what domain the text represents, they will not be able to use the appropriate register for the domain. The problem with domain and the NMT is that there may not be a sufficient amount of domain-specific training data, yet the NMTs are expected to perform well (Koehn 2020, 294). Koehn explains that, while NMT and SMT systems perform similarly in in-domain contexts, having applications for which each is better for, the out-of-domain performance of NMTs plummet while SMT systems' performance does not suffer significantly (ibid.). The most important factor that must be considered is that, today, MTs are used widely to translate natural language, while it has traditionally been considered that pre-editing requires modifying natural language to a CL. The specificity of the domain for which a controlled language has been designed can be very vague or very specific. That said, the accuracy of the translation must also suffer the vaguer the ruleset is. We should examine the possibility and the difficulties in making controlled language rulesets for the sample text types that are featured in this research: News texts, fictional texts, and scientific texts. We can first

evaluate, based only on subjective evaluation, that the order of language specificity, from most specific to least specific, is as follows: (1) scientific text, (2) news text, (3) fictional text. This hierarchy is based on the following analysis of the features of text:

Table 10 The bases for the evaluation of language specificity of the different text types.

Linguistic feature	Occurs in scientific texts	Occurs in news texts	Occurs in fictional text
creative use of language		x	x
non-traditional grammar			x
words created by the author	x		x
unconventional syntax		x	x

As we can see from table 10, Fictional texts can feature all four of the features for which rules cannot be set in a CL, news texts can feature two and scientific texts can feature one. Based on this evaluation, we can argue that a CL made for scientific texts can have a stricter ruleset than CLs made for the other text types. The language that is used in academia is, overall, very strict in grammar, lexicon, and syntax, which makes academic texts easier to translate in-domain than other text types that have more lenient rules in the aforementioned features. Designing a universal controlled language for fictional texts could be considered almost impossible. The translation of fictional texts is typically considered transcreation. Ji (2022) describes transcreation as follows:

> While some scholars pursue translation as an aesthetic, creativity-driven activity, others explore the practical functions or social utility of translation in fulfilling unmet needs. The term transcreation is used here to denote a purposeful integration of both approaches to translation to serve the needs of diverse people for information in an accessible, equitable, inclusive, creative and engaging way. (Ji 2022, 1)

When transcreating, the translator does not adhere strictly to what is in the source text, instead their intention being to modify the language so that the final product is cohesive and semantically sound. Therefore, applying a CL ruleset to fictional texts would defeat their purpose of being creative texts. Fictional texts do not follow any kind of specific structure, which means that even an NMT, whose entire training data consisted of fictional texts, would probably not be able to accurately translate fictional texts because of the factors presented in

table 1.6. For news texts which, for the purposes of this research, combine features of scientific and fictional texts can be made a universal CL that probably would not significantly hinder the quality of translation.

Pre-translation, such as the method proposed in Miyata and Fujita (2021) could result in better quality and efficiency in the short term. If a human translator translated parts of the text whose translation can be predicted to feature a case-inflection, we can circumvent the problem of the MT engine not being able to translate case-inflections. Miyata and Fujita's procedure was conducted by first examining the MT engines output and assigning an accuracy score on a five-tiered evaluation scale. If the translation quality is assessed as being of a desirable level, the evaluator can proceed to the final step, and if not, the evaluator should make small edits that, in their opinion, can have a positive effect on the translation quality. They will, then, assess, which version, either edited or the original, results in the best quality translation, if any, and proceed to the final step, in which they end the procedure and proceed with the best quality translation (Miyata and Fujita 2021, 3). This procedure does not seem to be any more efficient than the standard procedure of PoE, if even as efficient. The way this may be made to be productive than PoE is if the data is gathered in a database similarly to how TMs can be updated, and their data overwritten if the desired style of translation is different from previous ones. Although, this operation is almost entirely similar to the operation of a standard TM, which is why the operation may not be at all more efficient than standard PoE. After the conception of the first NMT systems, the possibility of incorporating neural network models into other linguistic tools has been studied. A pre-translation memory base that uses neural networks for pattern recognition following certain rules that inform the system, where the instance is applicable, could be a solution for solving such problems as the low translation accuracy of case-inflections. The aforementioned model is very similar to a standard NMT system that learns from the bilingual input it receives. The new concept would then be a set of rules that inform the NMT, which specific patterns in the source language correspond to specific patterns in the target language. The concept can be thought of as a controlled language ruleset for the NMT.

5.2 Means to make NMTs translate accurately with limited training data

The findings of this research do not provide sufficient knowledge to assess the correlation between the amount of training data and translation accuracy, because the NMT engines that

were the research subjects are proprietary and knowledge of the resources that they use are not available to the general public. We can, however, assume that the amount of training data, as well as the design of the NMT model have an effect on the accuracy.

The disparity between training data between language pairs can hinder the NMTs ability significantly. According to Koehn and Knowles' experiment, while an NMT needs an immense amount of training data to start working at all, it outperforms an SMT when the training data gets to approximately 15 million words and from there onward (Koehn and Knowles 2017, 4). This means that the threshold of switching from using an SMT to using an NMT is when the training data is at or exceeds 15 million words. This result alone does not account for specificity of usage, but assuming that the training data that is used is the same for the NMT and the SMT, the NMT produces improved quality. Also, judging from the graph that Koehn and Knowles (2017, 4) provide, While the ability of the NMT skyrockets in the early stages of adding training data, the progress slows down the more training data it receives, while the SMTs ability increases steadily with the addition of training data. We can, then, assume that Google's, DeepL's, Microsoft's and Amazon's NMT engines have all been trained with at least a 15 million Finnish word corpus to work as desired.

The most evident way to fix the problem of low accuracy due to an insufficient amount of training data seems to be system optimization either by designing new NMT architectures that perform better than the previous ones or designing tools that assist the NMT engine in the translation. These tools can be designed to provide assistance before or after the translation. These tools could be ones that automatically edit specific structures in the text to ensure that the MT output is accurate, but this method needs some sort of a CL to be effective and, as such, would work well only in a specific domain. The BLEU model could be modified in a way to make it broader and thus more resourceful in pre-translation settings. The BLEU model currently works in a way that evaluates the correspondence of the machine translated words to any prior string-pairs that are in the used knowledge base that it deems to match what is intended in the translation (Vashee, 2019). The system is not useful in measuring translation quality, instead only being useful when evaluating MT quality (ibid.). Despite this, the model can be a better fit for an infrastructure that is intended to provide options to a human translator. The model can be expected to work with a CL infrastructure, as it can be used drawing knowledge from any bilingual database. The complete infrastructure would need to feature a CL that is comprehensive enough for the intended use of the text. This means that, for translating a large variety of textual styles that appear in a large variety of

domains, a vast comprehensive CL would need to be constructed. The construction of such a vast database of language samples may seem impossible and, without specific resources and cooperation, it is safe to assume that it is not going to be possible. An outline of a philosophical conundrum can be devised from these expectations: can humanity cooperate in such a way to create a universal formal code of usage for any language. Creating a universal lingua franca has been tried with little success. An example of this is Esperanto, which is wholly human made. Creating a formal code for any one language is easier if a proper authority can be created to oversee the use of language in formal contexts. We run into another issue: Are there enough resources for such a trailblazing system. Universal codes can of course be written and introduced to translator education programs across the globe. Although, this way we would need to trust that the code will be used without further supervision. All of these problems can be circumvented with the use of technology in translation, such as the infrastructures theorized in this thesis. For the infrastructures to be useful, however, the translation landscape has to progress even more to utilize technology as it has done for decades now in leaps of varying lengths.

6 Conclucions

In this study, we aimed to determine if the four most popular MT engines: Google translate, Microsoft translator, Amazon translate, and DeepL translator can translate Finnish case-inflections accurately. The MT engines were given samples that each featured structures that, when translated, could be predicted to feature one of the nine Finnish case-inflections: partitive, essive, translative, inessive, elative, illative, adessive, ablative, and allative. The samples were gathered from news texts, a fictional book, and a scientific book, so that the results would cover the MT engines' performance in each of these three textual styles. This kind of research has not been conducted as of yet with the specific language pair, which means that the methods of examining the results were drawn from multiple different sources by mixing methods. When gathering the samples, it was observed that some structures, whose translations could be predicted to feature a case-inflection, were translated correctly 100 % of the time by the MT engines and others were not used at all by the MT engines. The latter issue may be due to the case suffixes' low representation in the training data, the MT engines being able to substitute the use of the case suffix in favor of an adposition that is more frequently used, or that the variety of English that has been used in the sample data does not feature structures that correspond to the Finnish case suffixes. Most probably the reason for their exclusion is some combination of these factors.

The samples were inputted to the four MT engines that were accessed via Phrase TMS and the translations were given a grade on a seven-level evaluation scale. The grades provided by the evaluation scale were then used to target cases with which one or more of the MT engines had issues with and the issues could further be evaluated with the support of the supporting material.

The research setting was not perfectly suitable for this purpose since the samples were gathered with DeepL translator, which gave it a significant advantage over the three other MT engines and the only variable that was consistent for all four MT engines was context, because, when gathering the samples for this study, DeepL had significantly more context when translating. On the other hand, the amount of context may have also been a detriment for DeepL, since NMT performance tends to be hindered when translating longer units, although this problem mainly pertains to sentence length.

In the initial inspection of the results, we can see that the MT engines fared better than initially expected with the MT engines having very little problem with the news-texts and scientific-texts sections. The main problem they had was with the fictional-texts section, which was in line with what was expected. Fictious texts can be written with very little regard to syntactical and grammatical rules, and can feature convoluted semantical relationships, which all constitute problems for an MT. Nonetheless, the MT engines had some problems in each section with the news-texts section producing only one issue that was middling in severity and the scientific-texts section producing multiple issues that were either low or very low in severity.

When investigating the correlation of the occurrence of specific case-inflections and the NMTs performance in translating the samples as predicted, it was determined that there was some correlation between the occurrence and translation accuracy, but the causality-relationship could not be determined, because the inner workings of the MT engines could not be inspected, and the correlative relationships were not as straightforward as expected. The most prominent issue with the frequency of occurrence could be observed in the data gathering phase, where DeepL did not output the abessive, the comitative and the instructive at all. The frequency that the abessive and the comitative occur in the grammatical archive and Parole corpus 0.2 % and 0.1 % respectively (VISK § 1227). The reason the instructive did not appear in the data gathering phase remains unknown, although it may also be due to low frequency of occurrence. The instructive appears 2.0 % of the time in the grammatical archive's corpus but only 0.3 % of the time in the Parole corpus (VISK § 1227). The instructive may be highly specialized in use, which could mean that it always occurs within certain structures and thus, if the MT engines' training data included those structures, they would be able to correctly use the instructive nearly 100 % of the time, which would be unproductive to research. There was enough evidence to suspect that the MT engine was either unable to produce the case-inflections, or it was always able to circumvent their use, which is why they were not included in the study. The deviations were observed in cases where the frequency of occurrence of the desired case suffix and the case suffix that the MTs used were separated by more than two percentage points in both datasets used as reference, and both case suffixes were plausible to use with minimal syntactic restructuring, which the MTs often failed to perform. Albeit there was only one such deviation within the set of translation outputs that were examined, which supports the argument the frequency of occurrence does predict translation accuracy, as was expected.

With further qualitative analysis of the problems, it was quickly observed that the case suffixes' frequency of occurrence may not have been the only issue contributing to inaccuracy. While sentence length was predicted to be a source of issues for the MT engines' accuracy, the samples did not seem to feature long enough sentences for this issue to appear. Instead, similar issues were observed with convoluted syntactical structures, which raises the question: is sentence length actually an issue for NMTs, or is it syntactical ambiguity? The most probable answer to this question is that both features cause issues and the subject should be studied to determine how the issues could be terminated. Mainly the fictional-texts section featured multiple issues of smaller proportions, such as hanging prepositions, that were featured in only one of the studied samples. Contrary to what was expected, metaphors and other idiomatic structures did not cause significant issues to the MT engines. While Google, Microsoft, and Amazon did not always translate idiomatic phrases accurately enough for such text to be publishable, they could certainly process that the phrase in question was idiomatic and translated according to the corresponding Finnish idiom.

Dissimilar syntactic structure caused by different levels of agglutination between the examined languages did not seem to cause great trouble for the MT engines, but in some cases, it seemed to be the only explanation for the inaccuracies. Unsurprisingly, this issue could also often be linked to the case-inflections' frequency of occurrence and the idiomaticity of the text, which were seen as issues throughout the research process. The problem presented as the SL featuring an adposition that does not correspond in the conventional manner to the TL case-inflection. The problem is aggravated if the desired case-inflection does not occur frequently enough in typical language use. In one presented instance, the desired case-inflection did not correspond in a conventional manner, the case-inflection had a, more frequently used, adposition counterpart, and the phrase was idiomatic, which made the use of the counterpart invalid. This resulted in each MT engine, other than DeepL, presenting a translation that was inaccurate in a different way. The observation of layered issues such as the one presented provides reasoning for further technological possibilities aimed at solving such issues. It has to also be considered that the problem affects only a small number of people and, as such, solving it may not be so significant as to merit to allocation of enough resources to resolve it. On the other hand, the model that could solve these issues could be used to solve other issues presented in translating between other language pairs, or it could be modified to be useful in solving other issues.

After evaluating the issues considering all the possible issues that the MT engines may have had, the possibility of pre-editing was considered as a solution for these problems. Pre-editing has not been extensively used in translation at any point in time and the most useful pre-editing instances have used a CL that has been designed for a specific domain. The possibility of universal CL of sorts was proposed with the issues of such design being outlined. While analyzing the features of the three textual styles being studied in this thesis, it was observed that the creation of a universal CL would be extremely difficult since fictional texts often feature unconventional syntactical and grammatical structures and also words that the author has simply fabricated. A sort of universal CL was deemed possible with given enough resources for creating it, and extensive cooperation between language authorities to ensure that the CL would be used widely enough. Without such resources, pre-editing cannot be seen as more productive than the currently prevailing technique of post-editing. The conception of tools to assist in pre-editing were also deemed to be not efficient enough to make pre-editing more efficient than post-editing.

This research subject would greatly benefit from further research conducted with slightly varying viewpoints. A potentially interesting future research subject would be to research the frequency of occurrence of Finnish case-inflections in a NMTs training data. This method would give further insight in to how the frequency of occurrence affects the NMTs performance and could be extremely useful in assessing if the problem can be resolved with some sort of tool. The study would need to be conducted with a computational model that automatically searches for case suffixes, because a NMT's training data are too vast for a human to efficiently conduct such a task. For this method, the access to a NMT's training data would also be needed, which would probably exclude the four NMT systems that were used in this research. The resources available for this study were insufficient for suggesting intricate solutions for the problems that were observed and future research could greatly benefit from evaluating the performance of an MT engine whose architecture is known. With this knowledge, more accurate assumptions of the causes of the presented problems could be drawn and, while we were able to conceptualize solutions for the problems, with more knowledge, the concepts could be more useable.

References

Literature

Angelone, Erik, Maureen Ehrensberger-Dow, and Gary Massey. 2020. The Bloomsbury Companion to Language Industry Studies. London: Bloomsbury Academic.

Boxer, Diana. 2002. Applying Sociolinguistics: Domains and Face-to-face Interactions. Philadephia: John Benjamins Publishing Company.

Corpas-Pastor, Gloria, and Isabel Duran-Muñoz, eds. 2018. Trends in e-tools and resources for translators and interpreters. Leiden: Brill.

Eberhard, David M., Gary F. Simons, and Charles D. Fenning (eds.). 2024. Ethonologue: Languages of the World. 27th ed. Dallas: SIL International. http://ethnologue.com.

Henisz-Dostert, Bozena, Ross R. Macdonald, and Michael Zarechnak. 1979. Machine Translation. The Hague: Mouton.

Huumo, Tuomas 2023. ”A Cognitive Linguistic account of the Finnish cases”. In Jaakola, Minna, and Tiina Onikki-Rantajääskö (eds.): The Finnish Case System: Cognitive Linguistic Perspectives. Vol. 11. Finnish Literature Society. http://www.jstor.org/stable/jj.8816156.

Ive, Julia, Aurélien Max, and François Yvon. 2018. “Reassessing the Proper Place of Man and Machine in Translation: a Pre-translation”. Machine Translation 32, no. 4: 279–308.

Jaakola, Minna, and Tiina Onikki-Rantajääskö, eds. 2023. The Finnish Case System: Cognitive Linguistic Perspectives. Vol. 11. Finnish Literature Society. http://www.jstor.org/stable/jj.8816156.

Ji, Meng (ed.). 2022. “Science and Art: Transcreation: Introduction to Special Section: Transcreation and Accessibility”. Leonard 55, no. 3: 289–290.

Kamprath, Christine, Eric Adolphson, Teruko Mitamura, and Eric Nyberg. 1998. “Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English”. ResearchGate.

Koehn, Philipp. 2020. Neural Machine Translation. Cambridge: Cambridge University Press.

Koehn, Philipp, and Rebecca Knowles. 2017. “Six Challenges for Neural Machine Translation”. Proceedings of the First Workshop on Neural Machine Translation: 28-39.

Laverdure, Richard P. 1983. “Dangling Participles, Hanging Prepositions, and Other Crimes Against the English Language”. The Army Lawyer: 25–29. HeinOnline Law Journal Library.

Marzouk, Shaimaa. 2021. “An In-depth Analysis of the Individual Impact of Controlled Languages Rules on Machine Translation Output: a Mixed-methods Approach”. Machine Translation 35: 167–203.

Miyata, Rei, and Atsushi Fujita. 2021. “Understanding Pre-editing for Black-box Neural Machine Translation”. Proceedings of the 16th Conference of the European chapter of the Association for Computational Linguistics: Main Volume: 1539–1550. ACL Anthology.

MQM. n.d. The MQM Error Typology. Accessed 7 March 2024. https://themqm.org/error-types-2/typology/

Nitzke, Jean, and Silvia Hansen-Schirra. 2021. A Short Guide to Post-editing. Berlin: Language Science Press.

Stephany, Ursula, and Maria D. Voeikova. 2009. Development of Nominal Inflection in First Language Acquisition: a Cross Linguistic Perspective. Berlin, New York: Mouton de Gruyter.

Tanwar, Ashwani, and Prasenjit Majumder. 2020. “Translating Morphologically Rich Indian Languages Under Zero-Resource Conditions”. Association for computing machinery. New York.

Vashee, Kirti. 2019. “Understanding Machine Translation Quality: BLEU Scores”. RWS Blog. https://www.rws.com/blog/understanding-mt-quality-bleu-scores/

VISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen, and Irja Alho 2004: Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura. Web version, accessed 6 March 2024. http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7.

Material sources

Carroll, Lewis. [1869] 2000. Alice’s adventures in Wonderland. Japan: Branden Books.

Dosanjh, Manjit. 2017. From Particle Physics to Medical Applications. Great Britain: Institute of Physics.

Guo, Owen. 2024. “The Chinese Millenials Shunning Lunar New Year Travel”. BBC News, 5 February 2024. Accessed 7 March 2024.

https://www.bbc.com/worklife/article/20240202-the-chinese-millennials-shunning-lunar-new-year-travel

Iovenko, Chris. 2024. “The Race to Build Climate-resilient Coral Reefs”. BBC News, 29 January 2024. Accessed 7 March 2024. https://www.bbc.com/future/article/20240123-the-race-to-build-climate-resilient-coral-reefs

Savage, Mark. 2024. ”The Highs and Lows of Grammys 2024 – and Why Taylor Swift Won Album of the Year”. BBC News, 5 February 2024. Accessed 7 March 2024. https://www.bbc.com/news/entertainment-arts-68200931

Sherriff, Lucy. 2024. “This Louisiana town moved to escape climate-linked disaster”. BBC News, 30 January 2024. Accessed 7 March 2024. https://www.bbc.com/future/article/20240130-this-louisiana-town-moved-to-escape-climate-disaster

Appendices

Appendix 1. Table of the results of the translations in the news-texts section

Case-inflection	Original sample	DeepL translation	Google translation	Microsoft translation	Amazon translation
Partitive	It cements her position as the one of the greatest songwriters of her era.	Se vahvistaa hänen asemansa yhtenä aikakautensa suurimmista lauluntekijöistä.	Se vahvistaa hänen asemaansa yhtenä aikakautensa suurimmista lauluntekijöistä.	Se vahvistaa hänen asemaansa yhtenä aikansa suurimmista lauluntekijöistä.	Se vahvistaa hänen asemansa yhtenä aikakautensa suurimmista lauluntekijöistä.
Essive	As a child, she relished simple rituals of the Lunar New Year.	Lapsena hän nautti kuun uudenvuoden yksinkertaisista rituaaleista.	Lapsena hän nautti yksinkertaisista kuun uudenvuoden rituaaleista.	Lapsena hän nautti kuun uudenvuoden yksinkertaisista rituaaleista.	Lapsena hän nautti yksinkertaisista kuun uudenvuoden rituaaleista.
Translative	I was trained as a conservationist and you do not interfere, you don't move things to new locations.	Minut on koulutettu luonnonsuojelijaksi, ja siihen ei puututa, eikä asioita siirretä uusiin paikkoihin.	Minut on koulutettu luonnonsuojelijaksi, etkä puutu, et siirrä tavaroita uusiin paikkoihin.	Minut on koulutettu luonnonsuojelijaksi, etkä puutu, et siirrä asioita uusiin paikkoihin.	Minut koulutettiin luonnonsuojelijaksi ja et puutu asiaan, et siirrä asioita uusiin paikkoihin.
Inessive	Lana Del Rey, calling her "a legend in her prime".	Lana Del Rey, jota hän kutsui "parhaassa iässään olevaksi legendaksi".	Lana Del Rey kutsui häntä "legendaksi parhaimmillaan".	Lana Del Rey, kutsuen häntä "legendaksi parhaimmillaan".	Lana Del Rey, kutsuu häntä ”legendaksi parhaimmillaan”.
Elative	Its aim is to discover new particles that would revolutionise physics and lead to a more complete understanding of how the Universe works.	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistaisivat fysiikan ja johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistavat fysiikan ja johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistaisivat fysiikan ja johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistaisivat fysiikkaa ja johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.
Illative	Its aim is to discover new particles that would revolutionise physics	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistaisivat fysiikan ja	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistavat fysiikan	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistaisivat fysiikan ja	Sen tavoitteena on löytää uusia hiukkasia, jotka mullistaisivat

	and lead to a more complete understanding of how the Universe works.	johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.	ja johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.	johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.	fysiikkaa ja johtaisivat täydellisempään ymmärrykseen maailmankaikkeuden toiminnasta.
Adessive	At the time of release, it was not her best-received album.	Julkaisuhetkellä se ei ollut hänen parhaiten vastaanotettu albuminsa.	Julkaisuhetkellä se ei ollut hänen parhaiten vastaanotettu albumi.	Julkaisuhetkellä se ei ollut hänen parhaiten vastaanotettu albuminsa.	Julkaisuhetkellä, se ei ollut hänen parhaiten vastaanotettu albuminsa.
Ablative	The final tier included residents who were displaced from the island prior to Hurricane Isaac.	Viimeiseen ryhmään kuuluivat asukkaat, jotka olivat joutuneet lähtemään saarelta ennen Isaac-hurrikaania.	Viimeinen taso sisälsi asukkaat, jotka joutuivat siirtymään saarelta ennen hurrikaani Isaacia.	Viimeinen taso sisälsi asukkaita, jotka joutuivat siirtymään saarelta ennen hurrikaani Isaacia.	Viimeinen taso sisälsi asukkaita, jotka siirtyivät saarelta ennen hurrikaani Isaacia.
Allative	The existence of a building block that gives all other particles in the Universe their form was predicted in 1964 by the British physicist, Peter Higgs	Brittiläinen fyysikko Peter Higgs ennusti vuonna 1964, että on olemassa rakennuspalikka, joka antaa kaikille muille maailmankaikkeuden hiukkasille niiden muodon.	Brittifyysikko Peter Higgs ennusti vuonna 1964 rakennuspalikan olemassaolon, joka antaa kaikille muille maailmankaikkeuden hiukkasille muotonsa.	Brittiläinen fyysikko Peter Higgs ennusti vuonna 1964 rakennuspalikan olemassaolon, joka antaa kaikille muille maailmankaikkeuden hiukkasille muodon;	Brittiläinen fyysikko Peter Higgs ennusti vuonna 1964 rakennuspalikan olemassaolon, joka antaa kaikille muille maailmankaikkeuden hiukkasille muodon

Appendix 2. Table of the results of the translations in the fictional-texts section

Case-inflection	Original sample	DeepL translation	Google translation	Microsoft translation	Amazon translation
Partitive	But do cats eat bats, I wonder?	Mutta syövätköhän kissat lepakoita?	Mutta syövätkö kissat lepakoita, ihmettelen?	Mutta syövätkö kissat lepakoita, ihmettelen?	Mutta syövätkö kissat lepakoita, ihmettelen?
Essive	burning with curiosity, she ran across the field after it, and was just in time to see it pop down a large rabbit-hole under the hedge.	Uteliaisuudesta palavana hän juoksi pellon poikki sen perässä ja ehti juuri ajoissa nähdä, kun se putosi pensasaidan alla olevaan suureen kaninkoloon.	palaten uteliaisuudesta, hän juoksi kentän poikki sen perässä ja oli juuri ajoissa näkemään sen putoavan suureen kaninkoloon aidan alla.	Uteliaisuudesta palaen hän juoksi pellon poikki sen perässä ja oli juuri ajoissa nähdäkseen sen ponnahtavan alas suuresta kaninkolosta pensasaidan alla.	Uteliaisuudesta polttaen hän juoksi kentän yli sen perään ja oli juuri ajoissa nähdäkseen sen putkahtavan suuresta kaninkolosta pensasaidan alla.
Translative	it was labeled "ORANGE MARMALADE," but to her great disappointment it was empty	siinä oli merkintä "ORANSSIMARMELADI", mutta hänen suureksi pettymyksekseen se oli tyhjä.	siinä oli otsikko "APPELSIINI MARMALADI", mutta hänen suureksi pettymyksensä se oli tyhjä	siinä oli merkintä "ORANGE MARMALADE", mutta hänen suureksi pettymyksekseen se oli tyhjä	se oli merkitty ”ORANGE MARMELADE”, mutta hänen suureksi pettymykseksi se oli tyhjä
Inessive	So she was considering, in her own mind	Joten hän harkitsi omassa mielessään	Joten hän harkitsi omassa mielessään	Joten hän harkitsi omassa mielessään	Joten hän harkitsi, omassa mielessään
Elative	the Rabbit actually took a watch out of its waistcoat-pocket.	Jänis todella otti kellon liivitaskustaan.	Kani itse asiassa otti kellon ulos liivin taskustaan.	Kani otti itse asiassa kellon liivitaskustaan.	Kani otti itse asiassa kellon vyötärötaskustaan.
Illative	she found herself falling down what seemed to be a very deep well.	hän huomasi putoavansa hyvin syvältä näyttävään kaivoon.	hän huomasi putoavansa hyvin syvältä kaivolta.	Hän huomasi putoavansa alas kaivosta, joka näytti olevan hyvin syvä.	Hän huomasi putoavansa alas, mikä näytti olevan hyvin syvä kaivo.
Adessive	yes, that's about the right distance—but then I wonder what Latitude or Longitude I've got to?	kyllä, se on suunnilleen oikea etäisyys - mutta mietin sitten, millä leveys- tai pituusasteella olen?	kyllä, se on suunnilleen oikea etäisyys, mutta sitten ihmettelen, mihin leveys- tai pituusasteeseen minulla on?	Joo, se on suunnilleen oikea etäisyys - mutta sitten ihmettelen, mikä leveys- tai pituusaste minulla on?	kyllä, se on suunnilleen oikea etäisyys - mutta sitten mietin, mihin leveysaste- tai pituusasteeseen minulla on?

Ablative	There seemed to be no use in waiting by the little door, so she went back to the table, half hoping she might find another key on it, or at any rate a book of rules for shutting people up like telescopes: this time she found a little bottle on it.	Pienen oven luona odottaminen näytti turhalta, joten hän palasi takaisin pöydän ääreen, puoliksi toivoen löytävänsä pöydältä toisen avaimen tai ainakin sääntökirjan, jossa oli sääntöjä, joiden mukaan ihmiset suljetaan kuin kaukoputket: tällä kertaa hän löysi pöydältä pienen pullon.	Ei näyttänyt olevan turhaa odottaa pienen oven vieressä, joten hän meni takaisin pöydän luo, puoliksi toivoen löytävänsä sieltä toisen avaimen tai joka tapauksessa sääntökirjan ihmisten sulkemiseksi kuin kaukoputket: tällä kertaa hän löysi siihen pieni pullo.	Pienen oven vieressä odottamisesta ei näyttänyt olevan mitään hyötyä, joten hän meni takaisin pöydän ääreen toivoen puoliksi löytävänsä siitä toisen avaimen tai ainakin sääntökirjan ihmisten sulkemiseksi kuin kaukoputket: tällä kertaa hän löysi siitä pienen pullon.	Pienen oven luona ei näyttänyt olevan mitään syytä odottaa, joten hän meni takaisin pöydän ääreen toivoen löytävänsä siitä toisen avaimen tai ainakin kirjan, jolla ihmiset suljetaan kaukoputkien tavoin; tällä kertaa hän löysi siitä pienen pullon.
Allative	she was walking hand in hand with Dinah, and was saying to her, very earnestly.	hän käveli käsi kädessä Dinahin kanssa ja sanoi tälle hyvin vakavasti.	hän käveli käsi kädessä Dinahin kanssa ja sanoi hänelle erittäin vakavasti.	hän käveli käsi kädessä Dinahin kanssa ja sanoi hänelle hyvin vakavasti.	Hän käveli käsi kädessä Diinan kanssa ja sanoi hänelle hyvin vilpittömästi.

Appendix 3. Table of the results of the translations in the scientific-texts section

Case-inflection	Original sample	DeepL translation	Google translation	Microsoft translation	Amazon translation
Partitive	On 4 July 2012, both the ATLAS and CMS collaborations announced that they had observed a new particle consistent with the Higgs boson.	4. heinäkuuta 2012 sekä ATLAS- että CMS-yhteisöt ilmoittivat havainneensa uuden hiukkasen, joka vastaa Higgsin bosonia.	4. heinäkuuta 2012 sekä ATLAS- että CMS-yhteistyö ilmoitti, että he olivat havainneet uuden Higgsin bosonin mukaisen hiukkasen.	4. heinäkuuta 2012 sekä ATLAS- että CMS-yhteistyö ilmoitti havainneensa uuden Higgsin bosonin kanssa yhdenmukaisen hiukkasen.	4. heinäkuuta 2012 sekä ATLAS- että CMS-yhteistyökumppanit ilmoittivat havainneensa uuden hiukkasen, joka oli yhdenmukainen Higgsin bosonin kanssa.
Essive	This can serve as a model for emerging multidisciplinary ventures in medical applications.	Tämä voi toimia mallina uusille monitieteisille hankkeille lääketieteellisissä sovelluksissa.	Tämä voi toimia mallina uusille monitieteisille hankkeille lääketieteen sovelluksissa.	Tämä voi toimia mallina uusille monitieteellisille hankkeille lääketieteellisissä sovelluksissa.	Tämä voi toimia mallina kehittyville monitieteisille hankkeille lääketieteellisissä sovelluksissa.
Translative	The work leading to the discovery—what The Economist lauded as ‘science’s great leap forward’ [2]—represented the culmination of decades of effort.	Löytöön johtanut työ - jota The Economist -lehti kutsui "tieteen suureksi harppaukseksi eteenpäin" [2] - oli vuosikymmenten ponnistelujen huipentuma.	Löytöyn johtanut työ - mitä The Economist kehui "tieteen suureksi harppaukseksi" [2] - edusti vuosikymmenien ponnistelujen huipentumaa.	Löytöön johtanut työ - jota The Economist ylisti "tieteen suureksi harppaukseksi eteenpäin" [2] - edusti vuosikymmenien ponnistelujen huipentumaa.	Löytöön johtanut työ - mitä The Economist kehui ”tieteen suureksi harppaukseksi eteenpäin” [2] - edusti vuosikymmenien ponnistelujen huipentumista.
Inessive	This was almost 50 years after the particle had first been predicted in theoretical calculations by Peter Higgs, Robert Brout and François Englert.	Tämä tapahtui lähes 50 vuotta sen jälkeen, kun Peter Higgs, Robert Brout ja François Englert olivat ennustaneet hiukkasen teoreettisissa laskelmissa.	Tämä tapahtui melkein 50 vuotta sen jälkeen, kun Peter Higgs, Robert Brout ja François Englert ennustivat ensin hiukkasen teoreettisissa laskelmissa.	Tämä tapahtui lähes 50 vuotta sen jälkeen, kun Peter Higgs, Robert Brout ja François Englert olivat ensimmäisen kerran ennustaneet hiukkasen teoreettisissa laskelmissa.	Tämä tapahtui lähes 50 vuotta sen jälkeen, kun hiukkanen olivat ensin ennustaneet teoreettisissa laskelmissa Peter Higgs, Robert Brout ja François Englert.
Elative	This mode of working has become second nature for particle	Tämä työskentelytapa on tullut luonnostaan hiukkasfyysikoille, jotka	Tämä työskentelytapa on tullut toiseksi luonnoksi hiukkasfyysikoille, jotka	Tästä työskentelytavasta on tullut toinen luonto hiukkasfyysikoille, jotka	Tästä työtavasta on tullut toinen luonne hiukkasfyysikoille, jotka

	physicists, who have learned to work collectively towards a common goal and who rely on consensus to take decisions.	ovat oppineet työskentelemään kollektiivisesti yhteisen päämäärän saavuttamiseksi ja jotka luottavat päätöksenteossaan konsensukseen.	ovat oppineet työskentelemään yhdessä yhteisen päämäärän eteen ja jotka luottavat konsensukseen tehdäkseen päätöksiä.	ovat oppineet työskentelemään yhdessä kohti yhteistä tavoitetta ja jotka luottavat yksimielisyyteen päätöksenteossa.	ovat oppineet työskentelemään kollektiivisesti kohti yhteistä päämäärää ja jotka luottavat konsensukseen päätöksenteossa.
Illative	It is clear that physics, and in particular particle physics, has made a major contribution to the development of instrumentation for biomedical research, diagnosis and therapy.	On selvää, että fysiikka ja erityisesti hiukkasfysiikka on vaikuttanut merkittävästi biolääketieteellisen tutkimuksen, diagnosoinnin ja hoidon instrumenttien kehittämiseen.	On selvää, että fysiikka ja erityisesti hiukkasfysiikka ovat vaikuttaneet merkittävästi biolääketieteellisen tutkimuksen, diagnoosin ja terapian instrumentoinnin kehittämiseen.	On selvää, että fysiikka ja erityisesti hiukkasfysiikka ovat edistäneet merkittävästi biolääketieteellisen tutkimuksen, diagnosoinnin ja hoidon instrumenttien kehittämistä.	On selvää, että fysiikka ja erityisesti hiukkasfysiikka ovat antaneet merkittävän panoksen biolääketieteellisen tutkimuksen, diagnoosin ja terapian instrumentoinnin kehittämiseen.
Adessive	CERN initiated a study to review the available technologies and determine what further developments would be needed to meet the requirements of this emerging treatment modality.	CERN käynnisti tutkimuksen, jonka tarkoituksena oli tarkastella saatavilla olevia tekniikoita ja määrittää, mitä lisäkehitystä tarvittaisiin tämän uuden hoitomuodon vaatimusten täyttämiseksi.	CERN käynnisti tutkimuksen tarkastellakseen käytettävissä olevia tekniikoita ja selvittääkseen, mitä lisäkehitystä tarvitaan tämän uuden hoitomuodon vaatimusten täyttämiseksi.	CERN käynnisti tutkimuksen, jossa tarkastellaan käytettävissä olevia tekniikoita ja määritetään, mitä lisäkehitystä tarvitaan tämän kehittyvän hoitomuodon vaatimusten täyttämiseksi.	CERN aloitti tutkimuksen käytettävissä olevien tekniikoiden tarkistamiseksi ja sen määrittämiseksi, mitä lisäkehitystä tarvitaan tämän nousevan hoitomenetelmän vaatimusten täyttämiseksi.
Ablative	These two independent detectors exploit different technologies, which is crucial for crosschecking and	Nämä kaksi toisistaan riippumatonta ilmaisinta hyödyntävät eri tekniikoita, mikä on ratkaisevan tärkeää uusien löydösten ristiintarkastuksen ja vahvistamisen kannalta.	Nämä kaksi itsenäistä ilmaisinta hyödyntävät eri tekniikoita, mikä on ratkaisevan tärkeää uusien löydösten tarkistamisessa ja vahvistamisessa.	Nämä kaksi riippumatonta ilmaisinta hyödyntävät erilaisia tekniikoita, mikä on ratkaisevan tärkeää ristiintarkistamiseksi ja uusien löytöjen vahvistamiseksi.	Nämä kaksi riippumatonta ilmaisimia hyödyntävät erilaisia tekniikoita, mikä on ratkaisevan tärkeää ristiintarkastuksessa ja uusien löytöjen vahvistamisessa.

	confirming any new discoveries.				
Allative	Physicists, engineers and computer scientists can share their knowledge and technologies, providing the medical community with first-hand information on the latest technical progress.	Fyysikot, insinöörit ja tietojenkäsittelytieteilijät voivat jakaa tietämystään ja teknologiaansa ja tarjota lääketieteelliselle yhteisölle ensikäden tietoa uusimmasta teknisestä kehityksestä.	Fyysikot, insinöörit ja tietojenkäsittelytieteilijät voivat jakaa tietämystään ja teknologiaansa tarjoamalla lääketieteen yhteisölle ensikäden tietoa uusimmasta teknisestä kehityksestä.	Fyysikot, insinöörit ja tietojenkäsittelytieteilijät voivat jakaa tietämystään ja teknologiaansa tarjoamalla lääketieteelliselle yhteisölle ensikäden tietoa viimeisimmästä teknisestä kehityksestä.	Fyysikot, insinöörit ja tietojenkäsittelytieteilijät voivat jakaa tietojaan ja tekniikoitaan tarjoamalla lääketieteelliselle yhteisölle ensikäden tietoa uusimmasta teknisestä kehityksestä.

Appendix 4. Finnish Summary

Tämän Pro Gradu -tutkielman tavoite oli tutkia, millaisia haasteita neljä konekäännintä kohtaa kääntäessään englanninkielisiä lauseita, joiden suomenkielisten käännösten voidaan odottaa sisältävän Suomen kielen sijamuotoja. Tarkastelun kohteena oli yksinomaan, osaavatko tutkimuksen kohteena olevat konekääntimet sijoittaa oikean sijamuodon oikeaan paikkaan ja kääntää ympäröivät asiat oikealla tavalla. Konekääntimien kykyä kääntää kieltä on tutkittu paljon erilaisten laatuparametrien avulla, mutta englannin kielestä suomen kielelle tapahtuvaa konekäännöstä ei ole aiemmin tutkittu sijamuotojen näkökulmasta.

Ensimmäinen konekäännin kehitettiin vuonna 1954, jonka jälkeen konekääntimien kehitys alkoi. Kehitys on tapahtunut aaltoillen vaihtelevan mielenkiinnon vuoksi. Ensimmäiset konekääntimet perustuivat sääntöpohjaiseen malliin. Vuonna 1988 IBM:n Peter Brown väitti, että tilastolliset mallit voivat toimia paremmin kuin sääntöpohjaiset mallit, jonka jälkeen statistisista malleista tuli käytetyimpiä konekäänninmalleja, kunnes neuroverkkopohjaiset mallit korvasivat ne 2010-luvun puolivälissä. Konekäännöksiä on käytetty pohjana jälkieditoinnille, jossa ihmiskääntäjä muokkaa konekääntimen tuottaman käännöksen kelvolliseksi loppukäyttöä varten. Jälkieditoinnin rinnalla esieditointi on ollut työtapa, jota on käytetty. Esieditoinnissa ihmiskääntäjä tai muu kieliasiantuntija muokkaa konekääntimelle syötettävää tekstiä ennen sen syöttämistä, jonka seurauksena teksti on sekä kieliasultaan että rakenteeltaan standardisoitua ja näin ollen konekäännin pystyy tuottamaan käännöksen, joka on valmis loppukäyttöä varten ilman, että ihmiskääntäjän on tarve muokata tekstiä jälkikäteen. Tässä tutkielmassa pohdin myös esieditoinnin mahdollisuutta ratkaista sijamuotoihin liittyvät ongelmat konekäännöksessä.

Neuroverkkokääntäminen ja jälkieditointi dominoivat nykyään kielialaa, koska kielillä, joista ja joihin suurin osa kääntämisestä tapahtuu, on saatavilla tarpeeksi aineistoa, jotta neuroverkkokääntimien toivotunlainen toiminta on mahdollista. Esieditointia on tutkittu myös neuroverkkokääntimien kanssa, mutta tutkimus ei ole ollut suosittua, koska näytöt esieditoinnin hyödyllisyydestä neuroverkkopohjaisia konekääntimiä käytettäessä ei ole ollut paljoa positiivista näyttöä. Tästä huolimatta esieditoinnin tutkimus neuroverkkokääntimien kontekstissa voi olla hyödyllistä ottaen huomioon neuroverkkokääntimien tunnetut kompastuskivet: niiden taipumus uhrata asiasisällön noudattaminen kielen sujuvuuden takaamiseksi, ja niiden tuottamien käännösten laadun heikentyminen, kun niihin syötetyn lauseen pituus ylittää 60 sanaa. Neuroverkkokääntimet kohtaavat myös merkittäviä haasteita,

kun käännettävä teksti ei ole tyylilajiltaan ja asiasisällöltään samanlaista kuin aineisto, jolla kääännin on ”opetettu”.

Suomen kielessä on 15 eri sijamuotoa, joita käytetään suffikseina muokkaamassa sanaa, johon ne on liitetty. Englannin kielessä näitä 15 sijamuotoa vastaa 19 rakennetta, jotka näyttäytyvät tavallisimmin prepositioina. Suomen kielen sijamuodot ovat: nominatiivi, akkusatiivi, genetiivi, partitiivi, essiivi, translatiivi, inessiivi, elatiivi, illatiivi, adessiivi, ablatiivi, allatiivi, abessiivi, komitatiivi ja instruktiivi. Kaikki suomen kielen sijamuodot eivät ole yhtä yleisiä kuin toiset. Sijamuotojen yleisyys kielenkäytössä vaikuttaa myös siihen, kuinka todennäköisesti niitä esiintyy aineistoissa, joilla konekääntimiä on ”opetettu”.

Tutkimusta varten keräsin osia uutisteksteistä, kaunokirjallisista ja tietokirjallisista teksteistä, jotka syötin DeepL:n konekääntimen läpi etsien sen tuottamista käännöksistä sijamuotoja. Esimerkkitekstit poimin BBC:n verkkosivulta, Lewis Carrollin kirjasta Alice’s adventures in Wonderland, ja Manjit Dosanjhin kirjasta From Particle Physics to Medical Applications. Arviointiin sisällytin suomen kielen sijamuodoista vain seitsemän kappaletta: partitiivin, essiivin, translatiivin, inessiivin, elatiivin, illatiivin, adessiivin, ablatiivin ja allatiivin. Tämän rajanvedon tein sen takia, että muiden suomen kielen sijamuotojen esiintyvyys on joko niin suuri, että konekääntimien voidaan olettaa automaattisesti käyttävän niitä oikein, tai niin pieni, että konekääntimien ei voida lainkaan olettaa osaavan käyttää niitä. Esimerkkien käännösten arviointiin kehitin löyhästi MQM laatumetriikkoihin pohjautuvan seitsenportaisen arviointiasteikon, joka perustui arviointikriteereihin. Arvosanat ja niitä vastaavat arviointikriteerit olivat seuraavat:

0. Kohdetekstissä ei esiinny sijamuotoa, eikä sen asiasisältö vastaa lähdetekstiä
1. Kohdetekstissä esiintyy sijamuoto, joka ei merkitykseltään vastaa lähdetekstissä esiintyvää rakennetta
2. Kohdetekstissä esiintyy toivottu sijamuoto, mutta se esiintyy väärässä paikassa, mikä muuttaa tekstin merkitystä.
3. Kohdetekstissä esiintyy sijamuoto, joka ei ole se, jota oli toivottu, mutta ei-toivotun sijamuodon käyttö ei merkittävästi muuta tekstin merkitystä.
4. Kohdetekstissä ei esiinny sijamuotoa, mutta se vastaa merkitykseltään lähdetekstiä

5. Kohdetekstissä esiintyy sijamuoto, joka ei ole se, jota oli toivottu, mutta ei-toivotun sijamuodon käyttö muuttaa tekstin merkitystä vain vähän.

6. Kohdeteksti sisältää toivotun sijamuodon ja vastaa asiasisällöltään lähdetekstiä.

Arviointiasteikko oli tarkoitettu tulosten helppoon vertailuun. Siitä huolimatta numeraalinen arviointiasteikko ei ole täydellinen tapa arvioida materiaalia, joka on vivahteikas, kuten kieli on. Tästä syystä tarkastelin aineistoa tarkemmin myös sanallisesti sen jälkeen, kun tulokset oli arvioitu numeraalisesti.

Käännettyjen tekstien arvioinnissa kävi ilmi, että konekääntimet käyttivät sijamuotoja useimmin oikein uutistekstejä ja tieteellisiä tekstejä käännettäessä. Kaunokirjallisia tekstejä käännettäessä konekääntimet käyttivät sijamuotoja merkittävästi heikommin kuin muita kahta tekstilajia käännettäessä. Uutistekstejä käännettäessä konekääntimien arviointien keskiarvot olivat arviointiasteikolla 5,66…–6, tieteellisiä tekstejä käännettäessä keskiarvot olivat 5,33…–5,77… ja kaunokirjallisia tekstejä käännettäessä keskiarvot olivat 3,66…–6. DeepL sai sekä uutistekstien että kaunokirjallisten tekstien käännöksistä täydet kuusi pistettä, sillä otokset oli etsitty materiaalista käyttäen DeepL-käännintä, joten oli ennalta odotettavissa, että DeepL pystyy käyttämään sijamuotoja oikein tässä tapauksessa. Tulosten arviointi aloitettiin tarkastelemalla Koehnin ja Knowlesin löydöstä, jonka mukaan neuroverkkokääntimillä on ongelmia pitkien lauseiden kääntämisessä (Koehn ja Knowles 2017, 1). Arvioinnin tuloksia ei kuitenkaan voitu tarkalleen vertailla Koehnin ja Knowlesin löydösten kanssa, sillä ainutkaan konekääntimille tässä tutkimuksessa syötetty lause ei ollut yli 60 sanaa pitkä. Koehn ja Knowles esittivät, että neuroverkkokääntimet kohtaavat ongelmia vasta lauseen pituuden ylittäessä 60 sanaa. Konekääntimille syötettyjen lauseiden enimmäispituudet olivat uutisteksteissä 26 sanaa, kaunokirjallisissa teksteissä 44 sanaa ja tieteellisissä teksteissä 51 sanaa.

Konekääntimet eivät kohdanneet merkittäviä ongelmia sana- tai lauserakenteissa, mutta joissakin tapauksissa sijamuotojen virheellinen käyttö johti lauseen sujuvuuden merkittävään heikkenemiseen. Ilmiötä tarkastellessa on kuitenkin otettava huomioon, että sujuvuuden heikkenemisen varsinaista syytä ei voida vahvistaa ja sijamuotojen puutteen ja sujuvuuden heikkouden suhde saattaa olla syy-seuraussuhde, puhtaasti vastaavuussuhde tai jotain niiden väliltä.

Kaunokirjallisten tekstien osiossa ablatiivi tuotti merkittäviä ongelmia konekääntimille, sillä niille syötetty lause, johon ablatiivia toivottiin, oli pitkä ja sisälsi prepositioita, jotka viittaavat alkuosassa esiintyvään substantiiviin. Mikään muu konekäännin kuin DeepL ei pystynyt tuottamaan toivottua ablatiivia käännökseen. Tässä tapauksessa ei kuitenkaan voida todeta, että konekääntimillä on ongelmia ablatiivin käytön kanssa, vaan ongelma saattaa johtua vain lauseen pituudesta ja monimutkaisista viittaussuhteista. Myös illatiivi tuotti konekääntimille ongelmia kaunokirjallisen tekstin käännöksessä, vaikka lause, jonka käännöksen voitiin olettaa sisältävän illatiivin, oli lyhyt. Tämän lisäksi illatiivi ei Ison Suomen Kieliopin aineiston mukaan ole harvinainen käytössä (VISK § 1227). Mikään konekääntimien tuottama lause ei sisältänyt illatiivia, vaan vaihtoehtoisen sijamuodon tai ei lainkaan sijamuotoa. Tässä tapauksessa mikään konekäännin ei tuottanut lähtötekstiä merkitykseltään vastaavaa käännöstä.

Ainoa sijamuoto, jonka käytössä konekääntimillä oli uutistekstiosiossa ongelmia, oli inessiivi. Tämä ilmiö viittaa siihen, että sijamuotojen käytön yleisyydellä ei ole vaikutusta siihen, pystyvätkö konekääntimet käyttämään niitä. Ison Suomen Kieliopin havaintojen mukaan inessiivi on tutkimuksen kohteena olevista sijamuodoista yleisin käytössä (VISK § 1227). Kaunokirjallisten tekstien osiossa voitiin havaita, että sijamuotojen yleisyydellä ja konekääntimien kyvyllä käyttää niitä oli jonkinlaista vastaavuutta. Kuitenkin yhdessä tapauksessa konekääntimet suosivat vähemmän yleisiä sijamuotoja yleisemmin käytetyn sijamuodon sijasta, joten vastaavuus saattaa olla täysin merkityksetöntä.

Konekääntimillä ei esiintynyt merkittäviä ongelmia suomen ja englannin kielten morfologisen rikkauden erojen kanssa. Kaunokirjallisten tekstien osiossa konekääntimet kuitenkin kohtasivat merkittävästi idiomaattisia rakenteita, joissa morfologisen rikkauden erot saattoivat olla yksi syy käännösten heikkoudelle. Esimerkiksi kaunokirjallisten tekstien osiossa illatiivin tapauksessa konekääntimien tulokset vaikuttavat siltä, että konekääntimille ei ole ollut saatavilla tarvittavaa kontekstia. Lause on lyhyt, joten ongelmana saattaa olla kontekstin puute. Lyhyet lauseet saattavat sisältää pronomineja tai muita sanoja, joilla on viittaussuhde johonkin toiseen sanaan. Ongelma siis esiintyy silloin, kun lauseessa esiintyy viittaava sana, mutta ei viittauksen kohteena olevaa sanaa. Lähtötekstissä morfeemi, joka merkittävimmin määrää sijamuodon käyttöä, esiintyy lauseen etuosassa, kun taas sijamuoto esiintyy lauseen loppuosassa. Lauseen keskellä esiintyy myös sivulause, joka vaikuttaa sekoittavan kääntimien toimintaa entisestään. Näin ollen tulokset vaikuttavat siltä, että kääntimet ovat ehtineet

”unohtamaan” sijamuodon määräävän morfeemin, siihen mennessä, kun ne ovat kääntämässä sanaa, jota sijamuoto määrittää.

Merkittävä ongelma konekääntimille vaikuttaa olevan se, että tekstit eivät ole tarpeeksi samankaltaista sen aineiston kanssa, jolla ne on koulutettu. Tämä seikka tulee erityisesti ilmi kaunokirjallisten tekstien osiossa, jossa kääntimillä oli eniten ongelmia. Tekstin esieditoinnin voidaan olettaa auttavan tähän ongelmaan. Erityisesti sijamuotojen käyttöön liittyviin ongelmiin voisi saada helpotusta, jos esieditoija muokkaisi tekstiä ennen konekäännöstä niin, että lähtökielen morfeemi, jonka voidaan ennustaa käännettävän sijapäätteeksi, ja sana, jota sijapääte tulee muokkaamaan, sijoitetaan lähekkäin käännettävässä lauseessa. Konekäännintutkimuksen tuloksia havainnoimalla voidaan huomata, että konekääntimillä oli merkittävästi ongelmia sijamuotojen käytössä, mikäli sana, jonka oli ennustettu kääntyvän sijapäätteeksi, ja sana, jota sijapääte kohdetekstissä määritti, olivat kovin erillään lähdetekstissä.

Ongelmana neuroverkkokääntimillä on myös niihin syötetyn koulutusaineiston käyttöalue. Neuroverkkokääntimet eivät toimi toivotulla tavalla, jos käännettävän tekstin käyttöalue ei vastaa koulutusaineiston käyttöaluetta. Koehnin mukaan neuroverkkokääntimet toimivat yksinkertaisempia tilastollisia kääntimiä heikommin, jos niiden kummankin koulutusaineisto on käyttöalueeltaan erilaista kuin käännettävä teksti (Koehn 2020, 294). Näin ollen kontrolloidun kielen käyttö neuroverkkokääntimien kanssa on hankalampaa kuin tilastollisien kääntimien kanssa, koska neuroverkkokääntimet vaativat moninkertaisen määrän koulutusaineistoa tilastollisiin kääntimiin nähden, jolloin niiden koulutukseen tulisi tuottaa niin laaja määrä tekstiä kontrolloitua kieltä käyttäen, että se ei ole mahdollista ilman erittäin mittavia resursseja ja niidenkin kanssa siihen kuluu erittäin paljon aikaa. Tästä huolimatta, jos neuroverkkokääntimelle antaa sääntöjä, joita sen tulee noudattaa, saadaan neuroverkkokäännin noudattamaan tiettyjä sääntöjä. Tämä olisi mahdollista toteuttaa esimerkiksi konekääntimen lisäosan avulla, joka esi- tai jälkikäsittelee tekstiä tiettyjen sääntöjen mukaan. Tällainen lisäosa vaatisi todennäköisesti jonkinlaisen kontrolloidun kielen ja voisi näin ollen olla hyödyllinen vain erittäin pienellä käyttöalueella. Näin ollen tämän tyyppisen järjestelmän luomisen voidaan ajatella olevan saatavilla olevien resurssien huonoa käyttöä, koska järjestelmää ei voida käyttää yleisesti kapean käyttöalueen vuoksi. Toisaalta tällaisen järjestelmän pohjalta voi olla helppoa luoda muita vastaavanlaisia järjestelmiä eri käyttöalueille.

Tutkielmassani esitetyn konekäännintutkimuksen merkittävimpinä ongelmina oli materiaalin keruun hankaluus ilman konekääntimen apua ja konekääntimien taipumus kiertää sijamuotojen käyttö. Materiaalin keruu tehtiin konekääntimen avustuksella, sillä ihmiskääntäjän kääntämiä tekstejä tulee analysoida tarkasti siltä varalta, että konekäännin pystyy helposti kiertämään sijapäätteen käytön esimerkiksi lauserakennetta muuttamalla. DeepL-kääntimen sisällyttäminen tutkimukseen sen jälkeen, kun sitä oli käytetty aineistonkeruussa ei ollut täysin tarkoituksenmukaista, vaikka DeepL ei suoriutunutkaan tutkimusosuudesta täydellisesti. Sijamuotojen esiintymisyleisyyden voidaan olettaa vaikuttaneen jälkimmäiseen ongelmaan. Tämä oli kuitenkin kierrettävissä, eivätkä konekääntimet pystyneet tasaisesti tuottamaan sujuvia käännöksiä käyttämättä sijapäätteitä. Konekääntimet suoriutuivat yleisellä tasolla paremmin kuin oli tutkimuksen alussa odotettu. Konekääntimien kyvyllä käyttää sijamuotoja toivotusti, ja sijamuotojen yleisyydellä todettiin olevan vastaavuussuhde, mutta syy-seuraussuhdetta ei voitu todeta tulosten epäjohdonmukaisuuden ja suljetun asetelman vuoksi.