Impact of sentence length on machine translation quality

Matić, Katharina

Master's thesis / Diplomski rad

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:684411

UNIVERSITY OF ZAGREB

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

DEPARTMENT OF ENGLISH

GRADUATE PROGRAMME

TRANSLATION

Katharina Matić

Impact of sentence length on machine translation quality

Master's thesis

Supervisor:

Nataša Pavlović, PhD

Zagreb, 2021

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA ANGLISTIKU

DIPLOMSKI STUDIJ

PREVODITELJSKI SMJER

Katharina Matić

Utjecaj duljine rečenice na kvalitetu strojnih prijevoda

Diplomski rad

Mentorica:

dr. sc. Nataša Pavlović

Zagreb, 2021.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Nataša Pavlović, for her patience and for agreeing to spend her time supervising this thesis. Without her guidance, which I am immensely grateful for, I would not have landed on this particular subject.

Additionally, I would like to thank Yousef Chanan, who graciously donated his time and technical knowledge in order to help me organise my data and compile the graphs used in this thesis. His patience knows no bounds.

I also thank my friends and colleagues, especially my boss Nina Tuđman Vuković, for continuously pushing me to write.

And finally, I must thank my parents for supporting me throughout my (longer than average) academic journey. Their love and encouragement have brought me where I am today.

Table of contents

Abstract

Neural machine translation (NMT), a relatively new language technology, has quickly taken over the place statistical machine translation models held as the most widely used machine translation method. But while it has many advantages in comparison to previously used methods, it also has its limitations, especially when it comes to translating long sentences.

This paper examines machine translations in both directions of translation for the Croatian-English language pair, to find out whether source text sentence length significantly influences machine translation quality. The central hypothesis to be tested is that machine translation quality drops proportionally to rising sentence length. The second hypothesis is that there is a "breaking point" in the number of tokens in source sentences after which there is a considerable drop in the translation quality.

Quality is defined as the ratio of the number of errors in translated sentences and the number of tokens in source sentences. The translations were generated using Google Translate in June 2019 and the manually compiled data set for human evaluation on which the study was conducted consists of three text types, one of each for each source language, as well as the translations of these texts.

Overall, based on the limited data set, the first hypothesis is confirmed, while confirming the second hypothesis would require a significantly larger data set. Nevertheless, the results indicate a tendency for the error score to plateau beyond the 40-token mark.

Key words

computational linguistics, neural machine translation, Google Translate, human quality assessment

Sažetak

Neuronsko strojno prevođenje (NMT) relativno je nova jezična tehnologija koja je vrlo brzo zauzela poziciju statističkih modela strojnog prevođenja kao najčešće korištena tehnika strojnog prevođenja. No iako ima mnogo prednosti u odnosu na prethodno korištene tehnike, ima i svoja ograničenja kada je riječ o prevođenju dugih rečenica.

U ovom se radu ispituju strojni prijevodi u oba smjera prevođenja za hrvatsko-engleski jezični par kako bi se utvrdilo utječe li duljina rečenice izvornog teksta znatno na kvalitetu strojnog prijevoda. Središnja hipoteza koju treba provjeriti jest da se kvaliteta strojnog prijevoda smanjuje usporedno s povećanjem duljine rečenice. Druga je hipoteza da postoji „prijelomna točka" u broju pojavnica u izvornim rečenicama nakon koje dolazi do znatnog pada kvalitete prijevoda.

Kvaliteta se definira kao omjer broja pogrešaka u prevedenim rečenicama i broja pojavnica u izvornim rečenicama. Prijevodi su generirani s pomoću Google prevoditelja u lipnju 2019., a ručno sastavljeni skup podataka za ljudsku procjenu na kojem je provedeno istraživanje sastoji se od triju vrsta tekstova na oba izvorna jezika te prijevoda tih tekstova.

Sveukupno je na temelju ograničenog skupa podataka prva hipoteza potvrđena, no potvrda druge hipoteze zahtijevala bi znatno veći skup podataka. Ipak, rezultati ukazuju na tendenciju da se udio pogrešaka stabilizira iznad duljine od 40 pojavnica.

Ključne riječi

računalna lingvistika, neuronsko strojno prevođenje, Google prevoditelj, ljudska procjena kvalitete strojnih prijevoda

1. Background

In a general sense, machine translation is described by Forcada (2010, p. 215) as "the translation, by means of a computer using suitable software, of a text written in the source language (SL) which produces another text in the target language (TL) which may be called its raw translation."

Neural machine translation (NMT) is a relatively new language technology, utilising neural networks for the purpose of automatically translating human language. The first study on the use of neural networks in machine translation, widely considered the beginning of neural machine translation, was published only eight years ago (Kalchbrenner and Blunsom, 2013), but despite that, this field has developed very rapidly.

Bahdanau et al. (2016, p. 1) define neural machine translation as an "approach to machine translation … [that] attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation." This is different from the previously widely used statistical approach, which was based on phrases as basic units of translation that were used in order to estimate translation probabilities from phrase pairs belonging to both the source and target language (Kalchbrenner, 2013, p. 1700).

The departure from previously used statistical models has had many advantages. Forcada (2010, p. 216) points out that disambiguation and differences in the structures various languages used to convey the same meaning were a frequent issue in ST systems, and the second issue in particular has seen great improvement with the development of NMT systems. This is confirmed by Bentivogli et al. (2016, p. 9), who state that "NMT has significantly pushed ahead the state of the art, especially in a language pair involving rich morphology prediction and significant word reordering."

While NMT has been found to provide higher quality translations for morphology-rich languages compared to statistical models (Bentivogli et al., 2016), a yet to be resolved issue, according to some research (Shi et al., 2021.), has been the translation of long sentences.

The topic of this paper centres around the relationship between sentence length and NMT quality. The aim is to examine the influence of sentence length on neural machine translation quality in order to point to existing problems in NMT.

Several authors have already addressed this issue. For instance, Bentivogli et al. point out that, according to available data, the quality of translation drops noticeably when translating longer sentences using NMT. Toral and Sanchez-Cartagena (2017, p. 1069) share similar findings, stating that PBMT (phrase-based statistical machine translation) outperforms NMT in sentences longer than 40 words, "with PBMT's performance remaining fairly stable while NMT's clearly decreases with sentence length."

Bahdanau et al. (2016, p. 1) further elaborate that most NMT models are based on the encoder-decoder principle, which operates by reading and encoding sentences into a vector of fixed length. This approach comes with its own issues, seeing as "a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector. This may make it difficult for the neural network to cope with long sentences, especially those that are longer than the sentences in the training corpus."

According to the data collected by Koehn and Knowles (2017, pp. 6-7) in their study conducted on the English-Spanish language pair, the statistical boundary of sentence length after which the quality of NMT compared to statistical models noticeably drops is approximately 60 tokens, and the difference in quality becomes even more pronounced after 80 tokens due to the fact that NMT has a tendency to shorten very long sentences, and therefore leave out some source content. Cho et al. (2014, p. 5) have also demonstrated that NMT systems underperform when translating long sentences compared to phrase-based systems.

However, this drop in quality is noted specifically in direct comparison to other systems, while the present study focuses on comparing the difference in translation quality based exclusively on varying sentence lengths within the same NMT system.

More specifically, this study explores the effects of source sentence length on neural machine translation quality, while focusing specifically on the language pair of Croatian and English. For the purposes of this paper, translation quality is defined as the ratio of the number of tokens in the source sentences and the number of errors in the translated sentences, through which an error score is calculated for each sentence.

The aim of this paper is to examine whether neural machine translation quality drops as the number of tokens in source sentences rises, and whether there is a "breaking point" in the number of tokens in a sentence after which translation quality drops more significantly. While there are precedents for such a breaking point being discovered in NMT, such as in Koehn and

Knowles (2017), this has not yet been done for the Croatian-English language pair, and more specifically, not done for this language pair only on NMT without comparisons to other types of machine translation.

Because NMT is such a rapidly developing field and Croatian is underresearched when it comes to machine translation, particularly NMT, there is a need for more detailed and goal-oriented research involving this language.

This subject is also relevant because practical NMT usage is rapidly increasing. For instance, EU institutions use the NMT-based eTranslation tool as a starting point template for translators and to generate placeholder translations on their websites. The system is trained on existing EU institution translations, which is useful for ensuring terminological consistency. Identifying and introducing a sentence length limit for writers of similar texts based on the results of this kind of analysis could serve as a guideline for writing texts that are meant to be subsequently translated. This would involve identifying a range of sentence lengths that is best suited for such texts, including a potential maximum sentence length.

In light of the above, the following two hypotheses are investigated:

H1: Neural machine translation quality drops with rising source sentence length.

H2: There is a breaking point in the number of source sentence tokens after which the drop in neural machine translation quality becomes more significant compared to shorter sentences.

The hypotheses are based on the results of previous work in this field, which appears to confirm that NMT underperforms when translating very long sentences compared to ST systems, and that there appears to be a noticeable and quantifiable drop in quality after a specific length compared to ST systems.

The differences in translation quality depending on direction of translation and text type were also compared and taken into consideration so as to determine the scope of their influence, if any at all, on final translation quality compared to sentence length itself, but this was a secondary consideration with no specific expectation attached.

2. Material and method

2.1. Material

The source data set was manually compiled and consisted of six subsets, half of which were originally written in English and the other half originally written in Croatian. Of these subsets, two comprised legal texts, two scientific texts, and two news articles, one for each language. This means that three subsets were compiled for each source language, one for each of the three text types analysed in this paper.

The source texts used in the data set were written on the same subjects within each text type category in order to make the source text pairs as similar in content as possible. The legal subsets contained excerpts from legal texts dealing with witness protection legislation, the news articles used for the analysis were written on the topic of Brexit, and the scientific texts consisted of excerpts from scientific papers about dementia.

Each subset contained roughly the same number of tokens, some slightly above and some slightly under 2000 tokens, and the overall total data set contained 12 580 tokens, while the overall total number of sentences was 506. More specifically, the source data set of Croatian texts contained a total of 265 sentences and consisted of a total of 6117 source tokens, while the source data set of English texts contained a total of 241 sentences and consisted of a total of 6463 source tokens. At the subset level, the news article, scientific and legal subsets in Croatian respectively contained 97, 88 and 80 sentences, while these subsets in English respectively contained 100, 93 and 48 sentences.

The sentence length in the cumulative data set ranged from 4 to 114 tokens, and the average sentence length overall was 24.9 tokens. More specifically, the average sentence length in the English source texts was 26.7 tokens and 23 in the Croatian ones. The average sentence length for the English source data set was therefore bigger compared to the Croatian, which was mainly due to the very long sentences in the English legal texts. The share of longer sentences containing 40 or more tokens in the cumulative data set was 10% (51 sentences, 30 in the English and 21 in the Croatian data set), of which 23 (4.5% of the total number, 17 in the English and 6 in the Croatian data set) contained 50 or more tokens.

As demonstrated, a variety of texts with differing average sentence lengths were chosen for the purpose of gathering enough data for each sentence length category. It should be noted that the final sample of sentence lengths was partially influenced by the availability of texts written on

the same topics for each of the analysed source languages while also having comparable sentence lengths.

The news article subsets contained the highest density of shorter sentences in both source languages, while the scientific texts were in the mid-range compared to the other two text types. The legal texts contained the longest sentences, with the one originally written in English containing the longest sentences overall. While the sentences in the Croatian legal text were, on average, longer than the sentences in the other two Croatian texts, their longest sentences were still noticeably shorter than the longest ones written in English, due to the difference in the general writing styles of English and Croatian legal texts.

Additionally, extremely long sentences are not generally common. As Pouget-Abadie (2014, p. 1) points out, "[t]raining on long sentences is difficult because few available training corpora include sufficiently many long sentences, and because the computational overhead of each update iteration in training is linearly correlated with the length of training sentences". As a result, most of the analysis was done on sentences under 60 tokens in length.

2.2. Method

The source texts were translated in June 2019 using Google Translate, a machine translation service offered by Google originally launched in 2006 as a statistical translation service. Google transitioned to NMT in November 2016 and has been using it since.

Both source and target texts were segmented into numbered sentences. Each of the source sentences was then marked with the number of tokens it contained, and each target sentence was analysed for errors according to previously outlined criteria and marked with the number of errors it contained. All of the error marking was done by the author of this paper.

For the purposes of this paper, very short sentences are defined as sentences up to 10 tokens in length (in line with Cho et al. 2014), short sentences as containing up to 20 tokens, medium sentences as containing between 21 and 39 tokens, and long sentences as sentences 40 tokens or longer, with very long sentences containing 50 tokens or more, seeing as most of the sentences from the texts analysed in this paper were below this length.

The broad criteria for determining translation errors used in this paper are based on the eight types defined by Simeon (2008, pp. 106-107), who lists them in the following manner: untranslated words, left out and added words, as well as lexical (semantic), orthographic,

morphosyntactic and stylistic errors, and errors regarding word order. More specifically, stylistic errors include atypical phrasing and unclear translations that could lead to confusion, while morphosyntactic errors include incorrect forms, subject-verb disagreement, incorrect use of articles and functional words, etc.

Depending on the nature of the errors discovered in the translations analysed in this work, this categorisation was subject to slight adaptations as appropriate for the requirements of this paper, although no significant changes were necessary, seeing as the kinds of errors found in the translations did not significantly deviate from these guidelines. The only significant addition to this categorisation was the category of semantic relation, mostly applicable to active and passive verb forms.

The perceived severity of the errors was not taken into account, as attempting to objectively measure the relative "weight" of errors in relation to each other is inherently subjective when taking into consideration the fact that the assessment of errors was carried out by the author of this paper. Furthermore, the nature of the errors in the translated texts is not the primary focus of this paper, but rather the number of their occurrences in relation to sentence length. Therefore, each occurrence was regarded in the same way for the purposes of data analysis. Pavlović (2017, p. 291) takes a similar approach.

The very act of deciding what constitutes an error or not is a subjective one, as pointed out by Ljubas (2017, p. 35), which is why the above criteria were used as a guideline according to which the assessment was made. However, at a more general level, even these guidelines are not definite enough to ensure completely objective and consistent error marking in the translated texts. Seeing as each case is somewhat unique and different, people might categorise or mark errors differently, even using the same guidelines.

In this particular case, all of the error marking was done by a single person, which is an advantage considering the relative internal consistency of the assessment, but also a disadvantage due to reduced objectivity, seeing as results based on a single person's assessment cannot be fully representative of a general trend without a number of other people participating to produce an average.

As with all human assessment, there is plenty of room for error. Nevertheless, the criteria provide a guideline strict enough to produce a data set that is marked for errors in a sufficiently consistent and justifiable manner.

Once the quantitative data on sentence length and errors was obtained, these numbers were used to calculate the error scores. The error score is defined as the ratio of target errors and source sentence length for each sentence in each subset, expressed as a percentage. The ratio of target errors to source sentence length was chosen, rather than the number of errors compared to the number of words in the translated sentences, because translated sentences usually do not contain the same number of words as the original sentences and in some cases they are differently segmented. In case of different segmentation, the dividing line for measuring the number of errors per sentence depends on the content of the source and target sentences. The resulting error scores were compiled into a total of nine separate graphs – one for each subset in each direction of translation, one for the total results for each direction of translation, and finally one graph containing all of the acquired data. These were then analysed and compared in order to either confirm or disprove the hypotheses of this paper; namely, whether translation quality drops in proportion to rising sentence length and whether there is a certain point in sentence length after which this drop in quality becomes more pronounced, with the drop in quality depicted in the graphs as a rise in error score.

The graphs were also examined for significant differences in order to gain insight into other potential factors that influence translation quality aside from sentence length, such as text type and direction of translation, seeing as the quality could also be affected by the number and types of texts the system has been trained on.

Additionally, all translation errors were colour-coded during the process according to the error categorisation laid out in this paper in order to compare the frequencies of each error type in each individual translated subset and see whether they showed any significant patterns. Another side question added to the analysis was whether there was any pattern or tendency regarding error distribution in machine translations of very long sentences, specifically whether more errors would accumulate in the first or second half of the sentence or rather be distributed evenly.

3. Results

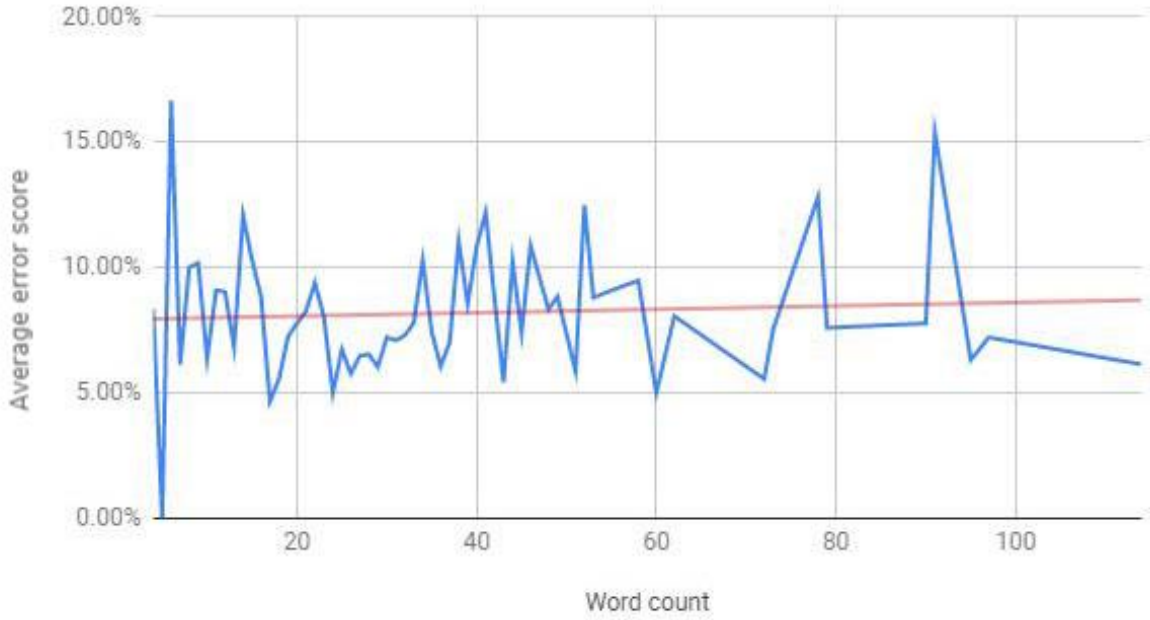

Figure 1: Graph of error scores in all analysed translations

The complete source data set for both source languages contained 12 580 tokens and the overall total number of sentences was 506. The sentence length ranged from 4 to 114 tokens, and the average sentence length overall was 24.9 tokens. The total number of translation errors was 970, making the average error score for the entire data set 7.5%.

The first hypothesis stated that NMT quality drops with rising source sentence length. The complete final graph in Figure 1 depicts the final cumulative results of the analysis, which shows a slight upward trend, confirming the first hypothesis laid out in this paper, albeit not very firmly. A much larger data set would produce more accurate results, as a bigger sample size with more examples of individual sentence lengths would provide a more faithful mirror of the general trend in machine translations for this language pair, and possibly others.

The second hypothesis, namely that there is a breaking point in the number of source sentence tokens after which the drop in NMT quality becomes more significant compared to shorter sentences, has not been confirmed by the results, as there is no visible point in sentence length at which the error score rises dramatically.

The following sections contain a more detailed breakdown of the results.

3.1. Croatian to English

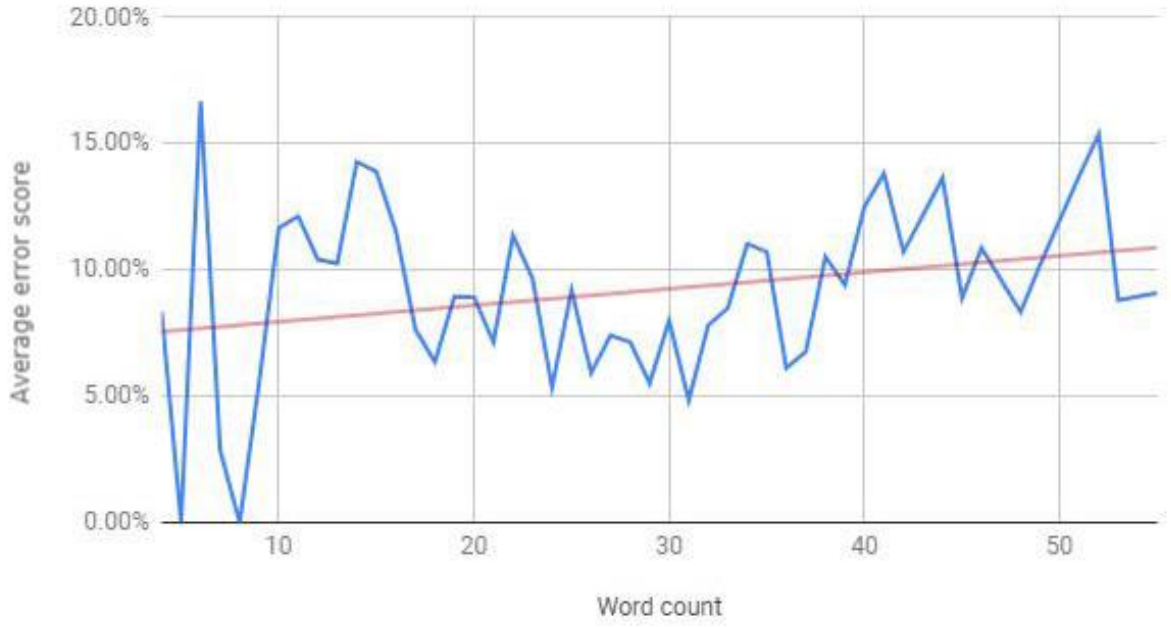

Figure 2: Graph of error scores in the English translation of the total Croatian data set

The source data set of Croatian texts contained a total of 265 sentences and 48 represented individual source sentence lengths. It consisted of a total of 6117 source tokens, while the translated data set had 542 translation errors. The most common type of translation error in this direction of translation were lexical errors.

Overall, the analysed machine translations from Croatian to English depicted in Figure 2 show an upward trend in their error scores as sentence length rises, confirming the first hypothesis. However, some spikes that are mostly absent in the subset graphs below appear in this cumulative graph, seeing as the error scores in two of the subset graphs below (Figures 4 and 5) lingered around 10%, while in one of them (Figure 3) this happened at 15%. The second hypothesis is not fully confirmed by the cumulative results for this subset, as there is no point on the graph where the error score begins to rise dramatically, but it does get slightly higher on average after the 40-token mark.

3.1.1. Croatian to English news articles

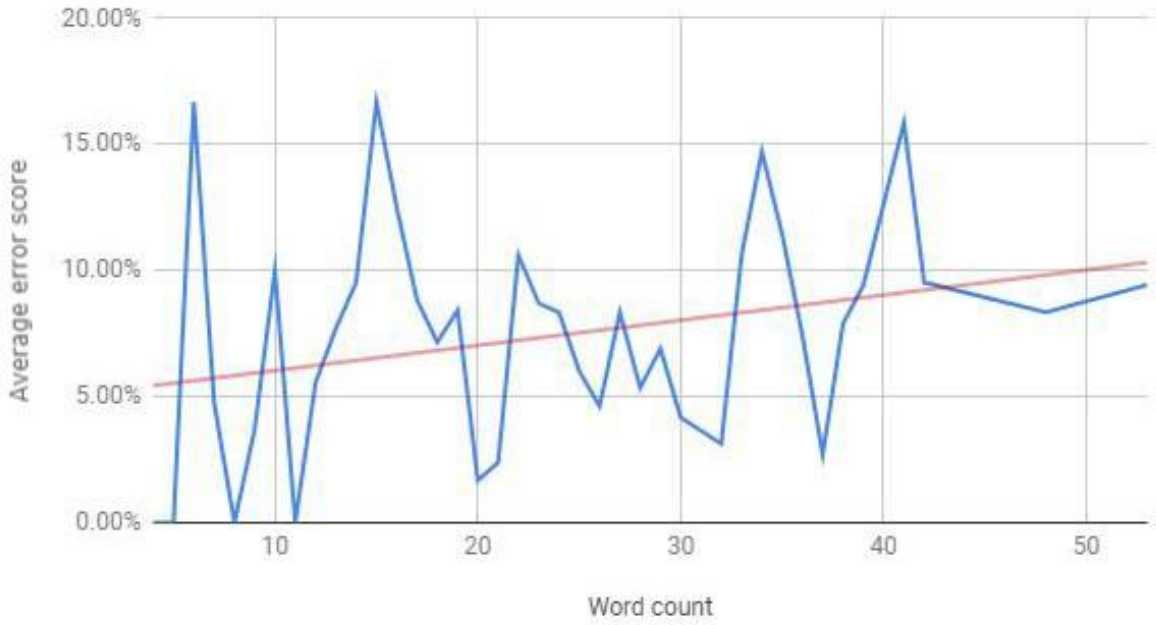

Figure 3: Graph of error scores in the English translation of the Croatian news article subset

The first subset in the data set of translations from Croatian to English was made up of excerpts from Croatian news articles about Brexit and contained a total of 97 sentences and 39 represented individual source sentence lengths. The results depicted in Figure 3 reflect the data gathered from the machine translation of this subset.

While most of the sentences had a fairly even distribution regarding length, there was a noticeably lower number of sentences longer than 50 tokens in this text type in particular, due to the nature of news articles, although this is also a general pattern that holds true for most of the cumulative data set for both directions of translation, which was a deciding factor in choosing the cut-off point for what is considered to be a very long sentence in the context of this paper.[1]

The results are in line with the first hypothesis, i.e. that error scores in translations rise proportionally to sentence length in the original text. There does not appear to be a point after which the error score rises significantly compared to the general trend, however, leaving the second hypothesis unconfirmed.

[1] In contrast, Cho et al. (2014, p. 7) define long sentences as starting at 30 words, but this kind of categorisation of long sentences did not seem appropriate for this paper based on the results.

Seeing as the subset is made up of news articles, very long sentences are not very common, so this sample alone does not necessarily reflect the breaking point we seek, if there is one. What is interesting, however, is that the error score does not appear to fluctuate as much as in shorter sentences after roughly the 40-token mark, but rather remains more stable compared to the rest of the graph. There is also an expected sharp spike in the error score on the left side of the graph, as even a single error takes up a big error score percentage in very short sentences.

There was a relatively high number of lexical errors in this translation compared to other error types as well as to the other texts in this direction of translation, notwithstanding occurrences of very rare words in the scientific subset. This could potentially be due to the fact that news articles are the least formulaic out of the text types analysed in this paper, especially compared to legal texts. However, the results for the translation from English to Croatian for this text type do not appear to have the same main issues as this translation, which is further discussed in section 3.2.1. Additionally, the fact that news articles contain shorter sentences on average means that such errors affect the intelligibility of their translations slightly more than in text types with a lower percentage of very short sentences. Semantic nuance is also lost in some cases, for instance in translations of the phrase *sporazum o razdruživanju*, which was in several instances translated as *disagreement agreement* by the NMT system.

3.1.2. Croatian to English scientific texts

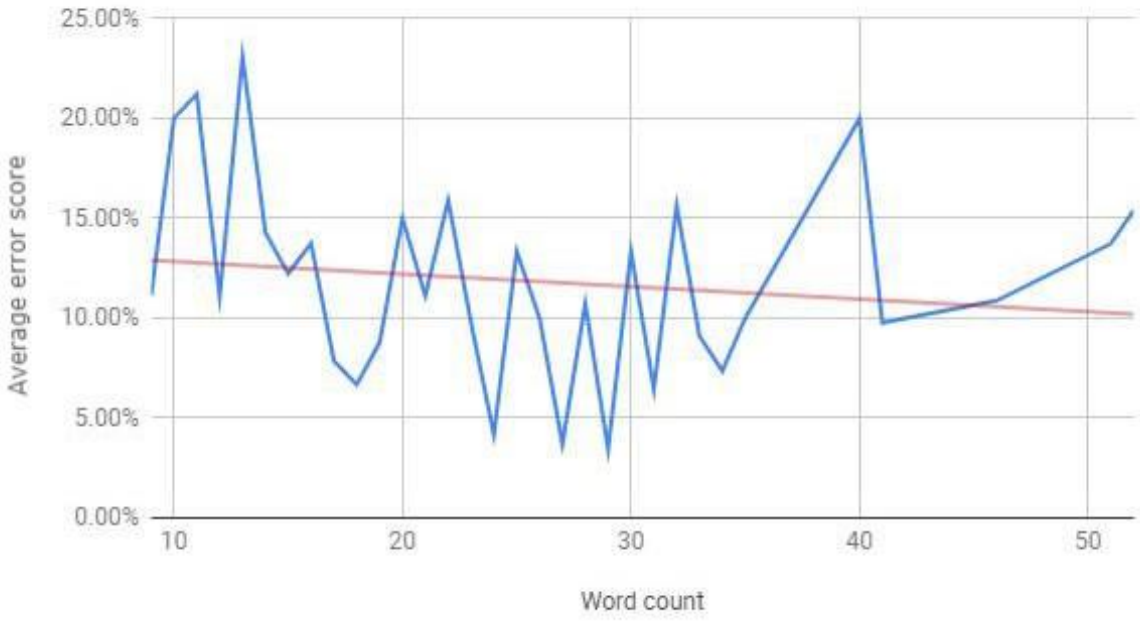

Figure 4: Graph of error scores in the English translation of the Croatian scientific subset

The second subset in the cumulative data set of translations from Croatian to English was made up of excerpts from a Croatian scientific text about dementia and contained a total of 88 sentences and 32 represented individual source sentence lengths.

The analysis of this translation shows very different results. The error score in the graph depicted in Figure 4 drops overall as sentence length rises, which is opposite to what was hypothesised. However, similarly to the previous graph, the error score once again appears to rise more consistently after the 40-token mark. This could be disputed due to the lower sample of sentences this length, but a lower sample would be more likely to have sharp spikes.

This subset was particularly problematic, as it contained acronyms based on source language terminology, as well as some rare words. It is therefore not surprising that this translation contained the highest number of lexical errors and untranslated words out of all the texts analysed in this paper.

For instance, the word *leukoaraioza* was mistranslated and in some cases left completely untranslated throughout the entirety of the text, despite the fact that the English translation shares the same Latin root and is very similar in form (*leucoaraiosis*). The same is true for the adjective *lakunaran*. This is evidently due to their low frequency, which is pointed out as a known shortcoming of NMT systems by Luong (2015, p. 1): "A significant weakness in conventional NMT systems is their inability to correctly translate very rare words: end-to-end NMTs tend to have relatively small vocabularies with a single unk symbol that represents every possible out-of-vocabulary (OOV) word."

The acronyms *TI* (standing for *tihi infarkt*) and *MR* (*magnetska rezonanca*) were also frequently used throughout the source text and left completely untranslated in the final translation. In the case of *TI* it could be argued that this particular acronym is not very frequent, but this does not apply in the case of *MR*, which should have been translated as either *MRI* or *MR imaging*, rather than just *MR*.

3.1.3. Croatian to English legal texts

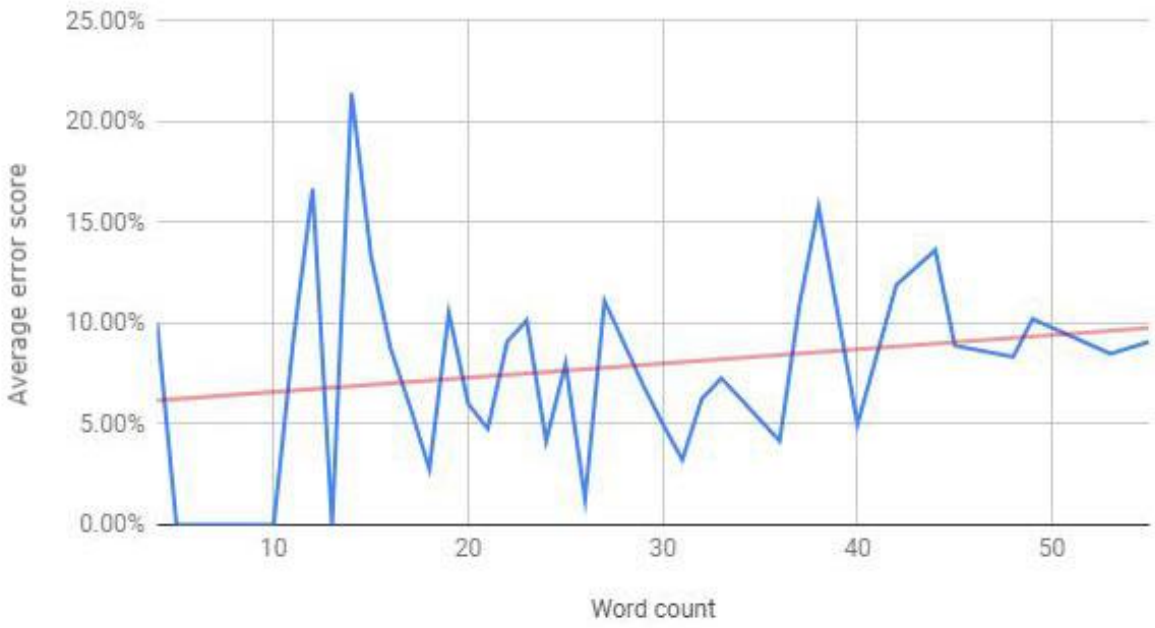

Figure 5: Graph of error scores in the English translation of the Croatian legal subset

The third subset in the cumulative data set of translations from Croatian to English was made up of excerpts from a Croatian legal text about witness protection and contained a total of 80 sentences and 37 represented individual source sentence lengths.

Figure 4 once again appears to confirm the first hypothesis, although the trend line is not as steep. And likewise, once again, there are fewer spikes after the 40-token mark.

The most common types of translation errors in this text were stylistic, morphosyntactic and lexical, in that order. The high occurrence of stylistic errors was surprising because of the formulaic nature of legal texts in general, as well as the sheer volume of such available texts that the NMT system could have been trained on, as it seemed that more stylistically uniform text types might produce more stylistically consistent translations.

For instance, the Croatian phrase *program zaštite* was frequently translated as *program of protection* instead of the more commonly used phrase *protection program*, although the correct translation of the phrase was used throughout the resulting text as well. The cause of this inconsistency is unclear, especially considering the fact that it was sometimes differently translated even in otherwise identical contexts.

3.2. English to Croatian

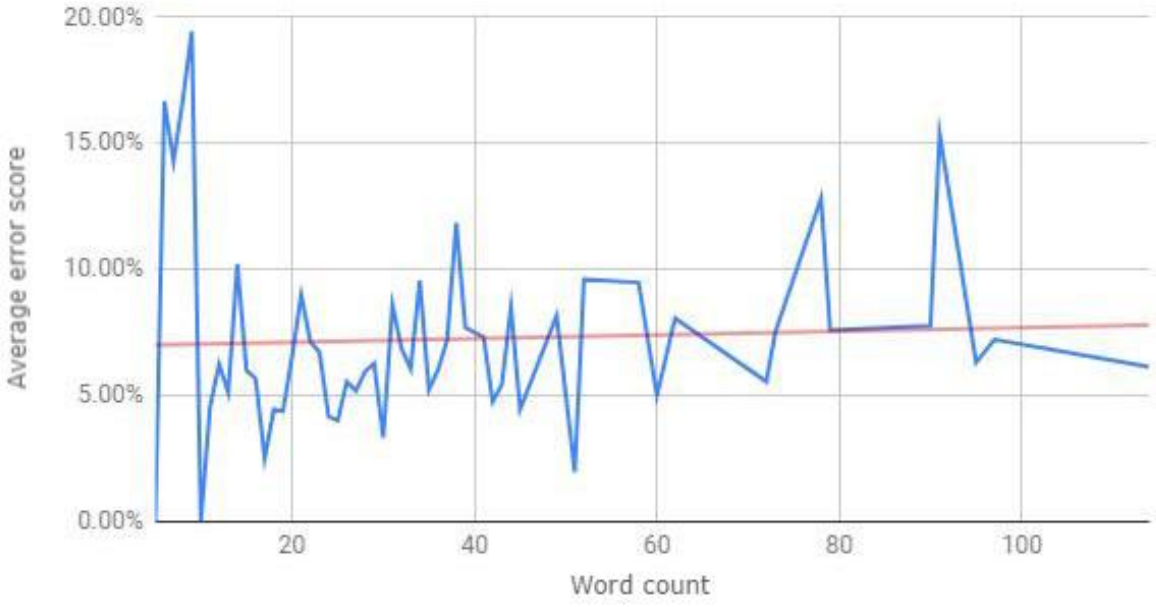

Figure 6: Graph of error scores in the Croatian translation of the total English data set

Once again, the final graph combining the results derived from all three machine translations from English to Croatian in Figure 6 shows a rising trend line, which is in line with the first hypothesis.

This data set consisted of a total of 6463 source tokens, while the translated data set had 428 translation errors. The most common type of translation error in this direction of translation were morphosyntactic errors, which is to be expected considering the richness of Croatian morphology, especially compared to English.

The data beyond the 40-token mark is less consistent than in some of the other examples, appearing to begin to plateau only around the 60-token mark. The sharp spikes following this spot on the graph are to be expected due to the underrepresentation of sentences containing that many tokens.

3.2.1. English to Croatian news articles

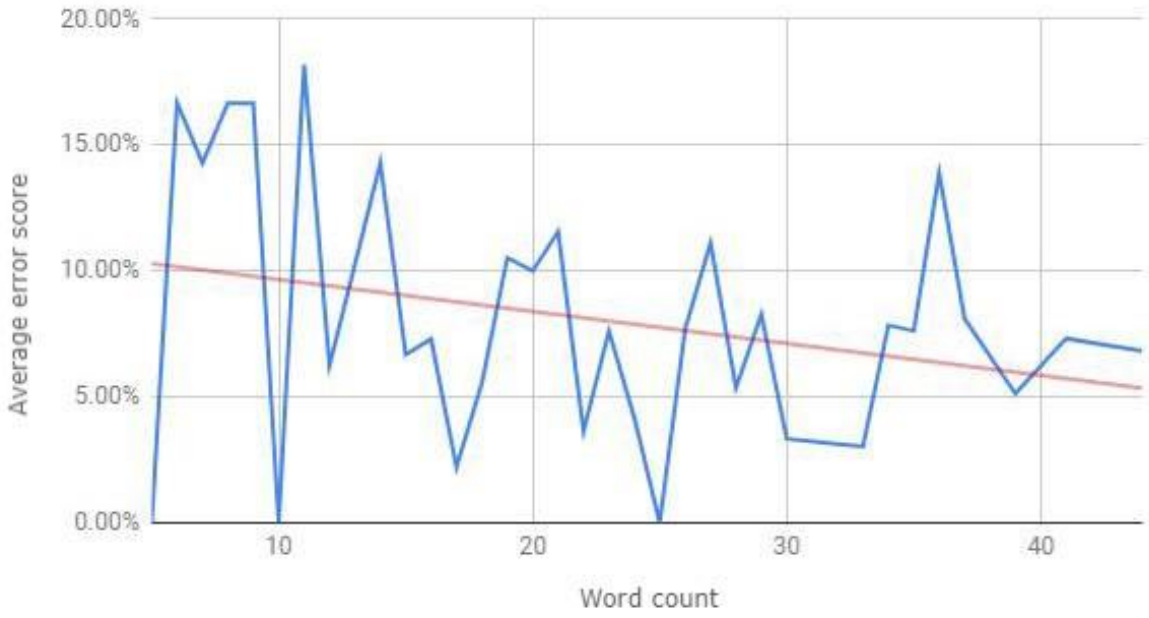

Figure 7: Graph of error scores in the Croatian translation of the English news article subset

The first subset in the cumulative data set of translations from English to Croatian was made up of excerpts from English news articles about Brexit and contained a total of 100 sentences and 36 represented individual source sentence lengths.

Unlike its Croatian pair subset, the results for the subset consisting of news articles originally written in English and translated into Croatian in Figure 7 show a downward trend in the error scores of its machine translation, with the highest error score appearing in the shortest sentences, which means that the results for this subset do not confirm the hypotheses.

Also, unlike the Croatian subset for this text type, lexical errors were not the most notable type of translation error, despite there being a moderately high amount of them. There was, however, quite a large number of morphosyntactic errors, which is to be expected due to the differences in complexity between English and Croatian morphology. It is a lot more difficult for the NMT system to accurately predict, for instance, the correct grammatical case or gender in Croatian than it is to translate this information into English, seeing as Croatian not only offers a lot more information as a starting point in translation due to its morphological richness, but also because English does not necessarily require all of this information to form a grammatically correct sentence with complete information structure. Agreement and grammatical gender in general were a frequent issue because the subset contained many mentions of a female prime minister,

and English does not morphologically contain information that would point to the correct gender to be used in the translation. Therefore, most cases like these simply defaulted to masculine translations. In contrast, acronyms such as *EU* agree with masculine verb forms in Croatian, while the machine translation produced a feminine verb in a sentence where *EU* was the subject. This acronym was an issue on another front as well, namely the complete lack of the declension necessary in Croatian in the translation.

3.2.2. English to Croatian scientific texts

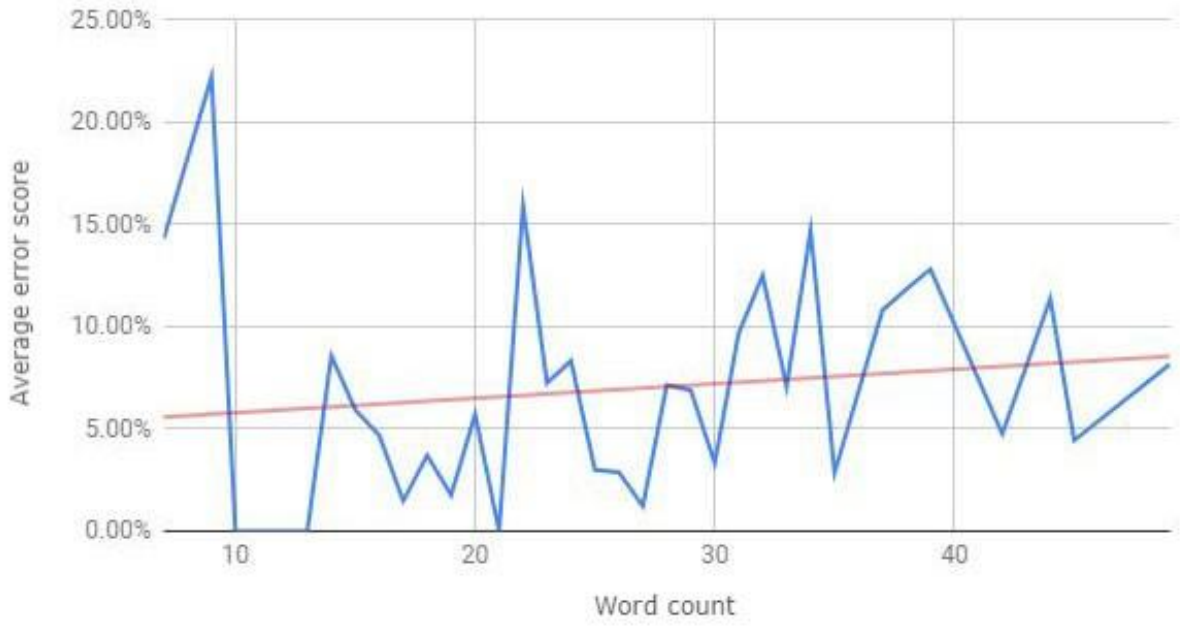

Figure 8: Graph of error scores in the Croatian translation of the English scientific subset

The second subset in the cumulative data set of translations from English to Croatian was made up of excerpts from an English scientific text about dementia and contained a total of 93 sentences and 34 represented individual source sentence lengths.

Figure 8 depicts the analysis of the machine translation of the English subset containing texts about dementia, which once again shows an upward trend, confirming the first hypothesis. The results after the 40-token mark are once again interesting, albeit not as obvious as in the examples where it somewhat plateaued. The spikes visible in the rest of the graph are still present beyond this point, but their baseline is higher compared to the spikes on the left side of the graph.

Morphosyntactic errors were once again the most common type of translation error in this particular subset. Similarly to its pair subset, the NMT system left acronyms untranslated. This was, purely coincidentally, appropriate for *VaD* (standing for *vascular dementia*), but not for *AD* (*Alzheimer's disease*). The appropriate acronym to use in Croatian would have been *AB* (*Alzheimerova bolest*).

3.2.3. English to Croatian legal texts

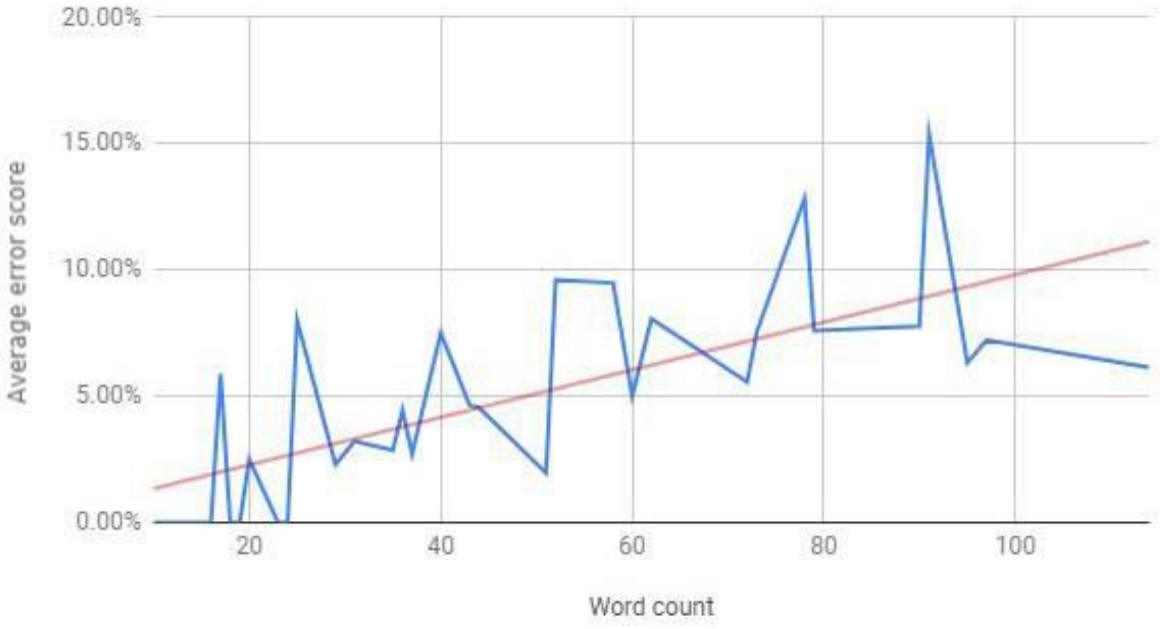

Figure 9: Graph of error scores in the Croatian translation of the English legal subset

The third subset in the cumulative data set of translations from English to Croatian was made up of excerpts from an English legal text about witness protection and contained a total of 48 sentences and 33 represented individual source sentence lengths.

Interestingly, this graph (Figure 9), representing the results for the text that contained the highest number of long sentences, as well as the longest sentences overall, matched the first hypothesis the best. The trend line is very steep and shows a noticeable rise in error score in direct correlation with sentence length.

Additionally, this is the only subset analysed in this paper that contained sentences longer than 60 tokens, even reaching over 100 tokens in length with its longest sentences. This means that all of the data in the final graph that refers to sentences longer than 60 tokens is in fact derived

from this single subset. Unlike the other subsets, the error score in this graph appears to plateau far beyond the 40-token mark, becoming relatively stable after a length of roughly 70 tokens.

There were many morphosyntactic errors in the translation of this subset as well, but the most common type of translation errors were stylistic errors, just like in the translation of its pair subset. It should be noted that for this text type it might not be the case that the text type itself results in a higher number of stylistic errors at a general linguistic level, but rather that this subset of linguistic usage has much stricter rules and therefore demands much higher precision and consistency in its translations, resulting in a higher perceived number of stylistic errors than usual.

3.3. Error distribution in translations of very long sentences

An additional question considered in the analysis was whether there was any discernible pattern or tendency regarding error distribution in very long sentences, in this case somewhat arbitrarily defined as sentences longer than 50 tokens, seeing as most of the sentences from the texts analysed in this paper were below 50 tokens in length.

The boxes shaded green in Table 1 represent the halves of long sentences which contained the highest number of errors, while the yellow boxes indicate an equal number of errors in both halves of individual sentences, with the total number of errors in each half highlighted at the bottom of the table.

Out of a total of 132 errors, 59 (44.7%) appeared in the first half, while 73 (55.3%) appeared in the second half of the analysed sentences. These results suggest that overall there are more errors in the second half of very long sentences translated using NMT.

The data set contained only 23 sentences longer than 50 tokens, further reinforcing this length as an adequate starting point for very long sentences. However, this also means that the sample itself may not be large enough to adequately represent a general trend in error distribution in sentences translated using NMT. Despite this, the results show a noticeable difference in the number of errors in each half and point to a possible tendency of NMT to accumulate errors in the second half of translated sentences. An analysis conducted at a bigger scale might uncover the extent of this tendency, if it does exist, and point to areas of necessary improvement in NMT systems.

words	errors	1st half	2nd half
51	1	0	1
51	1	1	0
51	7	4	3
52	5	3	2
52	8	3	5
53	3	2	1
53	6	2	4
53	5	4	1
55	5	3	2
58	4	1	3
58	7	0	7
60	3	3	0
62	5	1	4
72	4	1	3
73	3	1	2
73	8	4	4
78	10	8	2
79	6	3	3
90	7	2	5
91	14	6	8
95	6	1	5
97	7	2	5
114	7	4	3
		59	73

Table 1: Error distribution in translations of very long sentences

4. Discussion

The translation of the Croatian scientific text into English had the overall highest error score average, with a range between 10 and 15%, compared to the general range of 5-10% in other texts. The trend line in other texts barely passed the 10% mark – although, interestingly enough, the subset containing the longest sentences overall had a very steep trend line that crossed 10% past the 100-token mark and would seemingly have continued to extend past this percentage if the data set had contained even longer sentences. However, this is only conjecture, since there was no sample of longer sentences in the data set to confirm or deny it.

The graphs show different results for each direction of translation within the same text type – the translations of the subset of scientific text from English to Croatian showed a rise in error score as sentences grew longer, while the translations of the corresponding subset from Croatian to English showed a drop in error score in comparable conditions. Similarly, the translations of

the subset of news articles from English to Croatian showed a drop in error score with rising sentence length, while the translations of the corresponding subset from Croatian to English showed a rise in their error score.

The only text type that showed a rise in error score proportional to rising sentence length for both directions of translation were legal texts, which were also the subsets with the overall longest sentences, potentially suggesting that this trend becomes more stable for data sets containing very long sentences. Of course, this is not conclusive due to both the lacking size of the total data set, as well as the low variety of text types. While a larger sample might reveal whether sentence length or the text type itself is responsible, this particular text type appears to be inextricably linked with longer than average sentence length, and therefore it would be difficult to isolate the cause unless data sets from other text types with similar sentence lengths were also analysed under the same conditions.

Another factor not to be ignored is the high error score in very short sentences. Seeing as short sentences are also the most susceptible to showing sharp spikes in error scores, they have a definite influence on the slope of the resulting trend lines.

Furthermore, some words and phrases in the translated data set were incorrect in more than one way according to the criteria for error categorisation. For instance, the translation of the originally English scientific subset to Croatian contained several instances of the untranslated acronym *AD*. In this sense, it could have been categorised only as an untranslated word in all instances where it appeared. However, in many cases it was also morphologically incorrect, as it was missing inflectional morphemes, which are necessary to form grammatically correct sentences in Croatian. In such cases, this word was marked only with a morphosyntactic error rather than both morphosyntactic and untranslated, seeing as each instance was only a single token and therefore should not warrant more than a single error attached to it. This would also create a precedent for marking other words for errors in the same manner, which would make the method of error categorisation unclear and possibly inconsistent, as it can be difficult to not only assess, but also recognise how many errors should be assigned to a single token.

Additionally, marking a single word, especially one that is frequently reoccurring in a specific text, with more than one error would inflate the resulting error score in an unrealistic way and therefore negatively influence the results.

In the same vein, uninterrupted noun phrases that function as a whole and share the same morphological features were also counted as single errors rather than individually, because such errors jointly apply to the phrases, and varying phrase lengths might negatively influence the results as well. For instance, incorrect morphological case assignment in translations of very long noun phrases would significantly raise the error score of sentences that contain them, despite the fact that the error applies to the phrase as a whole. The opposite would be true for noun phrases containing a single word, even though the translation error itself is the same and applies the same way. This would mean that the focus would shift to issues related to translating sentences that contain very long noun phrases rather than sentence length in general, and would therefore not be an appropriate method of counting errors for the purposes of this paper.

In both directions of translation, the subsets consisting of legal texts had the lowest error score compared to subsets consisting of other text types in the same direction of translation. Out of all the data sets, the subset of legal texts translated from English to Croatian had both the lowest number of errors and error score overall, with only 130 translation errors in 2169 source tokens, making its total error score 5.99%. Its pair subset in the other direction of translation had 155 errors in 1976 tokens and an error score of 7.84%.

The highest error score was not as connected to text type, seeing as the text types of the subsets with the highest number and percentage of errors for each direction of translation belonged to different text types. In the data set of translations from English to Croatian, the subset with the highest error score was the translation of excerpts from English news articles about Brexit, with 152 errors in 2122 tokens and an error score of 7.21%. On the other hand, in the other direction of translation the subset with the highest error score was the one containing excerpts from a scientific text about dementia, with 213 errors in 1987 tokens and an error score of 10.72%. This was also the highest error score across all of the analysed subsets.

As for the remaining subsets, the scientific subset originally written in English contained 2172 tokens, while its Croatian translation contained 145 translation errors, making its error score 6.67%. The subset consisting of news articles originally written in Croatian contained 2154 tokens and its translation had 174 errors, making its error score 8.08%. Finally, the legal subset for the same direction of translation had 1976 tokens in the original text, 155 errors in the translation, and a 7.84% error score.

From the above stated information, it is evident that the data set of translations from Croatian to English as a whole had a higher error score than the data set from English into Croatian. It

contained a total of 542 translation errors in 6117 tokens (8.86%), compared to 428 translation errors in 6463 tokens (6.66%) in the other direction of translation.

It should also be noted that the word count in the original English data set (6463) was slightly higher than in the Croatian data set (6117) in order to compensate for the difference in word counts due to the frequent use of articles in the English language, and therefore create a more balanced linguistic sample.

Additionally, in order to truly have the variety in sentence lengths necessary for this analysis, certain compromises needed to be made in the compiling phase. Specifically, the legal texts written in both source languages contained the longest sentences overall, but only the one originally written in English contained sentences longer than 100 tokens. Instead of searching for similar texts containing sentence lengths more comparable to those in the Croatian legal subset, this text was chosen specifically for its representation of unusually long sentences that were more difficult to find in similar texts originally written in Croatian.

It was almost impossible to find two legal texts that would be parallel in both sentence length and topic, so this particular compromise had to be made in order to provide a relevant sample of very long sentences. However, this length difference inevitably leads to the issue that the longest sentences are only represented in the English legal text, but the cumulative results take into account both languages, and the general upward trend is still present in medium length sentences.

Interestingly, only the subsets consisting of excerpts from news articles each contained an entire untranslated sentence. This is in line with previously mentioned claims that NMT systems have a tendency to produce shorter translations, sometimes at the expense of informational content. Even more interesting is the fact that both sentences were not only very short – the Croatian original was only 9 tokens in length, while the English original sentence contained only 8 tokens – but they were also positioned at the very end of longer quotes consisting of several sentences enclosed in quotation marks. It may be possible that certain types of punctuation affect the completeness of NMT translations due to resulting segmentation errors, particularly if some segments are very short.

5. Conclusion

The results show that the data set of translations from Croatian to English cumulatively had a steeper rise in its error trend line than the data set of translations from English to Croatian, suggesting that machine translations from Croatian to English fit the first hypothesis of this paper better, seeing as the error score rises more noticeably with sentence length.

However, the English-Croatian translation data set contained much longer sentences than the Croatian-English data set due to the comparatively extreme length of some sentences in the English legal text used for the analysis. This means that the Croatian-English data set does not contain the data necessary to directly compare it to the English-Croatian data set, as their sentence length representations are not comparable.

Medium length sentences were far more common across all subsets, and therefore more accurately reflected the trend in error scores. Sharp spikes in the graphs were almost always the direct result of a single sentence length being underrepresented in the data set. Medium length sentences appear to be overall more balanced in this regard because their error-length ratios are not as extreme by default. The results suggest that medium length sentences, above 20 and below 40 tokens, are most likely the ideal length for texts written with the intention of being translated using NMT systems.

Additionally, the fact that only the news article translation subsets contained missing sentences might point to a possible need to pay closer attention to the influence of punctuation on content omission in translations produced by NMT systems, especially considering the fact that the other analysed text types did not contain such multi-sentence quotes. Accordingly, this might also mean that news articles and other texts that contain a higher number of direct quotes are more susceptible to content omission in translations created using NMT systems, which could be a useful factor for authors to consider when writing texts meant for machine translation.

Potential deviations were to be expected due to the limited scope of this work, i.e. the small sample size, the extremely fast development of the field discussed in this paper, and the fact that one of the selected languages is Croatian, which is a language that does not have the same level of NMT research and resources available as some bigger languages.

Due to the small sample size, the cumulative data set is not necessarily representative of general sentence length distribution in the selected text types and the results reflect this fact, seeing as

the data set contains only one sentence longer than one hundred tokens, and therefore the cumulative graph only reflects the error score of that single sentence for this sentence length.

One of the bigger issues in analysing the results using the criteria laid out in this paper was the fact that short sentences were far more likely to be interpreted as low-quality due to the fact that even a single error in a five-token sentence immediately results in a 20% error score. The severity of the errors in question could not be taken into account either, as this is far more difficult to assess objectively.

Text type did not appear to significantly influence the results regarding either hypothesis, although there are indicators that one of the text types analysed (particularly the English source legal subset) might be more in line with the hypotheses than the other two, which is mainly due to the difference in average sentence length among different text types.

Direction of translation did not appear to be a deciding factor either, seeing as the error score analysis for both directions of translation resulted in one dropping and two rising trend lines in their respective graphs.

Overall, the first hypothesis, namely that NMT quality drops proportionally to rising sentence length, has been confirmed. The same cannot be said of the second hypothesis, although there does appear to be a tendency for the error score to plateau beyond the 40-token mark, but confirming this would also require a much larger corpus with a much more diverse sample of sentence lengths, as well as examples of very long sentences across various text types and genres.

I hope that this work will help writers of texts meant to be used in machine translation and highlight existing problems in machine translations from and into Croatian, as well as the need for further development of digital resources for the Croatian language in order to increase their efficiency, seeing as it is necessary for the survival of modern languages to keep up with technological advancements.

References

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

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

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Forcada, M. L. (2010). Machine translation today. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 1, pp. 215-223). John Benjamins.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. *Proceedings of the 2013 conference on empirical methods in natural language processing* (1700-1709).

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Ljubas, S. (2017). Analiza pogrešaka u strojnim prijevodima sa švedskog na hrvatski. *Hieronymus – časopis za istraživanja prevođenja i terminologije*, *4*, 28-64.

Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

Pavlović, N. (2017). Strojno i konvencionalno prevođenje s engleskoga na hrvatski: usporedba pogrešaka. In D. Stolac & A. Vlastelić (Eds.), *Jezik kao predmet proučavanja i jezik kao predmet poučavanja* (pp. 279-295). Srednja Europa.

Pouget-Abadie, J., Bahdanau, D., Van Merrienboer, B., Cho, K., & Bengio, Y. (2014). Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *arXiv preprint arXiv:1409.1257*.

Shi, X., Huang, H., Jian, P., & Tang, Y. K. (2021). Improving neural machine translation with sentence alignment learning. *Neurocomputing*, *420*, 15-26. https://doi.org/10.1016/j.neucom.2020.05.104

Simeon, I. (2008). *Vrednovanje strojnoga prevođenja* [Unpublished doctoral dissertation]. Faculty of Humanities and Social Sciences in Zagreb.

Toral, A., & Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.