Ana Sabljak

SPEED-TO-QUALITY RATIO IN FULLY HUMAN TRANSLATION VS. POST-EDITING OF MACHINE TRANSLATION OUTPUT

Abstract

With an ever-growing amount of texts to be translated, often under very tight deadlines, there is growing pressure in the industry to use translation technology to speed up the translation process. More and more often, translators are called upon to post-edit machine translation (MT) output rather than translate a text. The aim of this study is to examine whether post-editing of MT output leads to greater speed and quality in translation. This study compares the speed and quality of fully human English-Croatian translations to those of post-edited MT output for three different text types: a novel, a news report, and a legal text. Quality is assessed along four parameters: morphosyntactic features, semantic features, style, and general impression. The MT system used for the purposes of the study is Google Translate.

1. Introduction

The quantity of texts that need to be translated is significantly increasing on a daily basis. A large number of novels are published every day, and many of them are translated into dozens, maybe even hundreds of languages. The same goes for news reports on all possible subjects, project documents, software and accompanying documentation, movies, TV commercials, TV series, websites, tourist brochures, cookbooks, user manuals for various household appliances, tools, equipment or electronic devices, etc. Due to the European Union's policy of multilingualism, according to which all official EU languages enjoy equal status (The Council Resolution of 21 Nov 2008), the workload faced by translators working at and for the various EU institutions is huge. Indeed, prior to Croatia's recent accession to the European Union, the entire EU legislation, the *acquis communautaire*, had to be translated into Croatian. Many Croatian legislative acts had to be translated into EU languages, mainly English, French and German,

pursuant to the EU institutions' "selective translation principle" (European Commission 2014), to prove that Croatian legislation had been harmonized with the *acquis.*

Considering the gigantic, almost never-ending amount of texts that are to be translated, it is only natural for translation providers to try to simplify the process and make translating faster and more efficient. This is precisely one of the main reasons why translators today use different tools that help them provide high quality translations in as little time as possible.

It is interesting to note that machine translation (MT), i.e. translation done entirely by a computer, was actually one of the earliest conceived applications of computers in general. However, completely automatized computer systems that would provide users with high quality, publishable translations of texts without any human intervention still do not exist. Nevertheless, computer science has been developing rapidly and today it is hard to imagine how the process of translating would even work without the technical support of computers and the tools translators use on a daily basis. Such translation, done by human translators who use different computer-aided translation tools (CAT tools), is called machine-aided translation or computer-aided translation. On the other hand, machine translation refers to the "automatic process of translating from one language into another using computers" (Tadić 2003: 162; my translation).

Having in mind this virtually inevitable symbiosis between human translators and computer software, I have decided to explore this topic and find out to what extent MT tools are helpful for the English-Croatian language pair and if they produce better results for some types of texts. Naturally, for the purposes of this paper I had to choose between different MT tools because it would be practically impossible to conduct research on all of them and include them in the scope of this study. I have chosen the one I consider the simplest, available to anyone with Internet access, and probably the most frequently used MT tool – Google Translate.

What is the quality of such machine translations? Does it vary from one type of text to another? Are MT systems helpful or distracting to professional translators? To what extent is the post-editing (PE) of MT output indeed faster

than conventional translation? After a brief overview of MT history and previous research in the area, this paper reports on the study conducted in order to provide answers to the above research questions.

2. Machine translation

2.1 MT history

Although the idea of translating natural languages by a machine dates back all the way to the 17th century, it became a reality 300 years later (Hutchins 1995: n.p.). In the 1930s, patents for mechanical devices that would function as a bilingual dictionary were issued to Georges Artsrouni (France) and to Petr Troyanskii (Russia) (Hutchins 1997: n.p.). However, their ideas remained unrevealed until the late 1950s. In the meantime, theoretical designs were drawn in the Weaver memorandum in 1949, which stimulated much interest and research in the field. During the 1950s, the idea of mechanical translation gained momentum, especially in the US, where the Government provided significant funding. Optimistic atmosphere was in place and fully automated high quality machine translation was envisioned for the near future. Nevertheless, the first systems used the word-for-word method, and the development of formal grammar was still suffering from teething problems, so the enthusiastic bubble burst soon thereafter. The ALPAC Report in 1966 gave a strong blow to the development of MT technology in the US. It claimed that MT was slow, more inaccurate than, and twice as expensive as human translation, and that "there was no immediate prospect of MT producing useful translations of general scientific texts" (Arnold et al. 2001: 13). The ramifications of the report were considerable, in that the US Government brought the funding in the field of MT to an end, and it drastically reduced the number of experts interested in further development of MT technology. However, smaller groups of researchers still continued their work, which resulted in the emergence of the SYSTRAN system, a Russian-English MT system, which is actually an improved version of the earliest MT systems presented at the Georgetown University in the 1950s. SYSTRAN was used by organizations such as the United States Air Force and NASA, and its later development led to the occurrence of French-English and Italian-English

versions. Another example of a successful MT system occurred in Canada in 1976. METEO system was successfully translating weather reports from English into French and vice versa, due to limited vocabulary and specific sentence structure (Arnold et al. 2001: 12-15). During the 1980s and 1990s, there was a significant development of MT systems around the globe (the USA, Japan, Germany). In the 1990s, in the course of further development of computer technology and the Internet, statistical MT systems appeared. Toward the end of the 20th century, it became easy to develop and share electronic corpora and online terminology databases, and first online MT systems appeared. Nowadays they have acquired the status of commercial products used every day all over the world by professional translators, as well as personal users that need to translate texts for all possible purposes. (Arnold et al. 2001; Hutchins 1995)

2.2 Types of MT systems

There are different classifications of MT systems, depending on the method they use. Some authors divide them into rule-based systems, statistical systems and example-based systems; others, such as Hutchins (1995: n.p.), distinguish between rule-based and corpus-based systems. Tadić (2003: 37) divides MT systems into rule-based and empirical systems, dividing the latter further into statistical systems and example-based systems. Many authors divide MT systems into two basic groups, rule-based and statistical systems, which is the classification I decided to follow, since it seemed detailed enough for the purposes of this paper.

Rule-based systems use sets of rules that enable the translation process from a source language into a target language. The philosophy that lies behind rule-based MT systems can be compared to the process of acquiring a foreign language for humans. Learning a foreign language is a complex process. In simplified terms, we can say that acquiring a foreign language consists of learning the vocabulary of a certain language and learning grammatical rules according to which words can be combined to form word phrases, clauses, and other grammatical structures. Rule-based MT systems use the same methodology. They are based on the process of substituting source language words by words belonging to the target language, according to extensive

bilingual dictionaries. Subsequently, the word order is rearranged according to the rules of the target language. However, every human language is an extremely complex system, there is often no word-for-word correspondence between languages, and almost no rule comes without an exception, while in many cases the exceptions are numerous. Including all those rules and exceptions in a computer program used for translating, i.e. mimicking the human process of acquiring language when trying to "teach" a machine a language, is not only an extremely complicated and almost impossible task, but also one that results in translations that are of unsatisfying quality. (Tadić 2003: 37-38; Google 2013)

On the other hand, statistical MT systems are based on data. They work in such a manner that they use bilingual or multilingual parallel corpora (i.e. large quantities of texts originally written by humans in a specific source language, and of aligned translations of those texts, called parallel texts for a specific language pair), trying to recognize translation equivalents and to choose the most appropriate one among them. During this process, statistically-based MT systems use various statistical methods, such as probability, in order to find translation equivalents calculated from available parallel corpora (Tadić 2002: 37-39). The result of this process, the generated translation, is then offered to the user. It is logical, therefore, that the larger the number of texts in the source language (SL) and the target language (TL), the greater the quality of the offered translation. Having in mind that not all languages are represented evenly according to the number of texts available to MT systems, the quality of translation varies from one language pair to another (Google 2013).

Statistical MT systems were initially word-based, but this proved to be an inadequate approach since, as already mentioned, there is no word-for-word correspondence between any two languages in the world. More specifically, the problems might occur with homonyms (*bank* in English could either mean a 'financial institution' or a 'river bank'), fixed phrases, collocations and idioms (e.g. the correct Croatian translation of the English idiom "still waters run deep" is "tiha voda brijege dere", but a word-for-word MT might come up with a nonsensical translation such as "ipak vodama pokrenuti duboko"). Later, phrase-based systems have proved to be capable of producing translations of greater

quality. Some other specific ways of expression, such as sarcasm, might also be hard for a machine to recognize or translate adequately. In addition, translating cultural references is a highly demanding task from a translator's point of view. There are several methods translators can resort to in order to address the issue, all of which require a great deal of knowledge of the world, as well as knowledge of the language, combined with a thorough understanding of both source and target cultures. Computers are not yet ready to deal with such complex tasks successfully. A good example are references to popular characters in the source culture, which are not widely known in the target culture. For instance, some American movies and TV shows are abundant with references to *Gilligan's Island*, a sitcom from the 1960s that was never broadcast in Croatia. Any human translator familiar with the Croatian culture is aware of that fact and would probably substitute the *Gilligan* reference with a more familiar one, and therefore more appropriate for the Croatian viewer. In contrast, a computer cannot be aware of such circumstances, and would probably leave the *Gilligan* reference as it is, rendering its connotation completely unintelligible.

Moreover, machine translations may be grammatically correct, yet seem clumsy to a native speaker of the target language, who can easily detect that there is something wrong. Important factors that human translators can benefit from are common sense, intuition, knowledge of the world, knowledge of culture, and their own life experience, which computers lack.

2.3 Google Translate

Google Translate (GT) is a statistically-based empirical MT system. As already mentioned, the quality of translations it generates depends on the size of parallel corpora available for a certain language pair. This is a multilingual MT system, able to translate between any of the 80 languages currently supported by GT, using English as an intermediary language for most combinations (Bellos 2011). Google Translate is "trained on the Europarl Corpus (Koehn 2002), the DGT Multilingual Translation Memory (European Commission Directorate-General for Translation 2007) and the United Nations ODS corpus (United Nations Official Document System 2006)" (Uszkoreit et al. 2010: 1104).

Google Translate has improved significantly over the last few years. More and more features have been developed, such as the possibility for users to edit translations the program offers, and to do this in a very simple manner: should a user not be satisfied with a certain word or phrase within GT translation, it is enough to click on that word and the program offers other possible solutions. Other features include the *Listen* button, which allows users to hear the actual pronunciation of certain words or whole texts; GT also functions as an online dictionary: by writing a single word inside the box, the user is provided with a range of possible meanings. The quality of translations improves as more and more texts are available for GT to use when finding established patterns.

Today there is also a Google Translate application for smartphones. Not only does it translate written texts, but also spoken phrases. This function is still somewhat limited; nevertheless, it is becoming more and more sophisticated.

3. Previous research in the area

Considering the relevance of the topic for professional translators, it comes as no surprise that there has been a lot of research exploring the quality of machine translation and comparing fully human translation (FHT) to post-editing of MT output.

An interesting study was conducted by Koponen (2010), in which the author set out to establish criteria for assessing translation quality, focusing on the accuracy of semantic content in translation. After proposing an error classification, the author compared fully human translation to machine translation, as well as two different types of MT systems to one another. The rule-based system used in this study was a demo by Sunda Systems Oy, whereas the statistical system was Google Translate. Three different types of texts were translated from English into Finnish. The results showed that, for the statistical system, the most common error was omitting the relation between two concepts, which often resulted in an unconnected list of concepts. On the other hand, the most common errors with the rule-based system were mistranslating an individual concept and mistaken relations between the concepts, which resulted in more convincing sentences at first glance, but they actually turned

out to be misleading. While human translators also made mistakes in terms of adding and omitting concepts, there was a notable difference with respect to MT systems – as a rule, concepts added by human translators were in a way related to the source text, which was not the case with machine translations (Koponen 2010). This was particularly interesting for the present study because it points to the above mentioned "handicap" of MT systems in terms of their lack of knowledge of the world. MT systems seem to lack the fine line and logic telling them that a certain concept is an intruder in a specific context. In contrast, human translators would grasp this immediately and they need to make no conscious effort in order to avoid such mistakes.

In another related study, Calude (2003) used the SYSTRAN system to compare the performance of machine translation systems with respect to four different text types: extracts from technical instructions, a short story, a news report, and a magazine article. All texts were translated from German into English and, as in Koponen's study, the author also classified errors. Of the four text types, MT proved to be completely useless for translating the short story extract, and the quality was so poor that the author said that not a single TL sentence made sense. In contrast, translating the technical set of instructions, MT produced the least number of linguistic errors, and the most frequent errors referred to polysemous words translated outside the proper context, or structural differences between the two languages, e.g. faulty word order, wrong prepositions or literally translated phrasal verbs. Similar errors, but greater in number were observed in machine translations of newspaper and magazine article extracts. The results of this study largely coincide with the results I obtained in my own research, as will be seen in section 6 below.

Fiederer and O'Brien (2009) conducted a study with the aim to establish whether MT output necessarily was of lower quality than human translation. They used a user guide in English as a source text; the target language was German, and the MT system used was IBM WebSphere. The study involved 11 evaluators, or "raters", who rated the source sentences, the translated sentences and the post-edited sentences, taking account of three parameters: clarity, accuracy, and style. The results were in favour of machine translated, post-edited output when it came to the accuracy and clarity parameters, but the human translators were

more successful according to the style parameter. The parameters were similar to the ones used in the present study, as will be seen in section 6.3 below.

4. Aims and hypotheses

The primary aim of this research was to check to what extent MT systems, more specifically Google Translate, are useful tools in the translation industry. In particular, I wanted to compare fully human translation with post-editing of MT output in terms of speed and translation quality.

I expected MT to speed up the process, but also to perhaps distract translators to some extent, inducing them to overlook or even introduce some mistakes into the final product. My first hypothesis was therefore that post-editing of MT would be faster than fully human translation, but of inferior quality.

Furthermore, I wanted to compare different text types in terms of their suitability for use with Google Translate. My hypothesis in this respect was that texts tending to use predictable language would lend themselves more easily to machine translation and post-editing translation process than texts that use language in a more creative way.

Regarding the given parameters, I expected the fully human translations to get higher scores on the morphosyntactic level for Text 1 (novel), having in mind the complexity of the Croatian morphology as opposed to the English morphology. Due to the specific features each of the three text types usually displays, I expected to find the biggest difference between FHT and the post-editing of MT output on the morphosyntactic level for the fictional text, and the smallest difference for the legislative text. On the semantic level, I expected the results to be in favour of the FHT for the fictional text, and for the other two texts I expected better results in favour of the post-editing of MT output. As for the style parameter, I expected the FHT to be more successful in the fictional text, and the post-editing of MT output to be more successful in the legal text, since the specific style of legal texts is much easier to follow, provided that such texts are highly represented in GT corpora. For the same reasons, the same results were expected for the general impression parameter.

5. Methodology

In order to test my hypotheses, I conducted a study consisting of two stages. The first stage involved an experiment in which one group of subjects produced a fully human translation of three different types of text, while another group post-edited MT output (produced by GT) of the same three texts. The time required for completing the task was measured for each text. The second stage consisted of evaluating both sets of translations by a group of evaluators. In continuation, I elaborate on the choice of texts, the research participants and the evaluation method.

5.1 Choice of texts

Three different texts of approximately 100 words each were selected for the experiment. While choosing the texts, account was taken of several factors – the appropriateness of the topic of each text (if the text was unbiased, free of political views, appropriate from the ethical point of view), the corresponding level of vocabulary for the purposes of this study, and whether the texts contained an appropriate variety of nouns, verbs, tenses etc. The first text was a fragment of a novel written by Kurt Vonnegut, *Slaughterhouse-Five or The Children's Crusade*. The second text was an excerpt of a news report about the nuclear catastrophe in Fukushima, Japan, from *The Guardian*, which had wide international media coverage, and the language seemed to meet the abovementioned criteria. *The Guardian* is a respectable newspaper, so I considered its website to be a legitimate source of possible texts for the research. The third text was a part of the EU Directive 2010/64/EU of the European Parliament and of the Council of 20 October 2010 on the right to interpretation and translation in criminal proceedings. I chose this text because, as previously mentioned, today there is a high demand for translation of such texts from English into Croatian. All three texts were originally written in English, and the study involved their translation into Croatian.

Having in mind the way Google Translate works, I expected the results of my research to show that this tool could be very helpful when translating texts abundant with fixed phrases that are always translated in the same manner. In

other words, GT was expected to be more suitable for texts in which the vocabulary, or rather terminology, is more restricted. On the other hand, texts belonging to fiction are often richer in vocabulary and figures of speech, and the style of the author is often discernible. Such complex and sophisticated linguistic characteristics, combined with the fact that such texts, as a rule, belong to very broad semantic domains, were the reason I did not expect GT to prove very helpful in the case of the excerpt from a novel. As for news reports, their linguistic features quite depend on the type of newspaper. I chose an objective report that did not display the author's personal style, and used frequent vocabulary. For this reason, I expected the results to be in favour of GT; that is, I expected them to show that GT was helpful when translating news reports of this kind.

5.2 Participants

The participants in the experiment were students at the end of their first year of the Graduate Study Programme in English, Translation Track, at the Faculty of Humanities and Social Sciences, University of Zagreb. All of them were native speakers of Croatian. I divided them into two groups, the Translators (11 students) and the Post-Editors (10 students). The Translators were given three pieces of paper with the source texts on top of each page, and were asked to translate each text in continuation, making a note of the time they started, as well as the time they completed each translation. A digital clock was projected on the wall so every student could see it. Once they decided a particular translation was finished, they were asked not to go back and revise it further. The Post-Editors were given GT-generated translations on top of each page and were asked to revise each, i.e. to rewrite the translation, changing only what they thought should be changed in the process, and nothing else. They were also asked to make a note of starting and end times for each text. The Post-Editors had access to the source texts, but were instructed to revise the GT-generated translations rather than translate from scratch. The participants were allowed to use dictionaries, but did not use computers. Clearly, this is an important limitation of the study, which will be further discussed in section 8.1.

5.3 Assessment method

In the second phase of the research, three Evaluators were asked to score the translations from the first stage of the research. All of the Evaluators met the following criteria:

1. They were native speakers of Croatian;
2. They had recently graduated from the English Department's Translation Track;
3. They had a high level of competence in English and in translation (all of them had graduated with honours);
4. They had had some professional experience.

The Evaluators were asked to score the translations according to four different parameters: (1) morphosyntactic features, (2) semantic features, (3) style, and (4) the general impression. Translations were coded in such a way that the Evaluators did not know which translations were fully human translations, and which were post-edited Google translations. Naturally, strict confidentiality was observed, and the names of the participants were never disclosed at any stage of the research, or in this paper.

After receiving all the assessments, I could commence processing the results. Using Microsoft Excel, averages were calculated for both groups, first taking account of the time variable for all three texts together, and then for each text separately. Each translation was then compared in terms of the scores it received from the Evaluators, taking account of each parameter separately. Average scores were calculated for each translation, and for each of the four parameters. The obtained results were then studied in detail, in order to draw some conclusions. Some of these tables are provided in the Appendices.

In the next section I will present the results I obtained, grouped according to various criteria.

6. Findings

6.1 Translators vs. Post-Editors – the time variable

The overall results showed that the Post-Editors, on average, needed 17 seconds more to complete their task than the Translators. This might not seem to be a notable difference at first glance. However, a more detailed insight revealed that the difference in time needed to complete the task between the Translators and the Post-Editors varied drastically across texts: the most notable difference was observed in connection to the news report, where the Post-Editors needed a whole minute and 8 seconds more to complete their task (Fig. 1). On the other hand, in the case of the legal text, the Post-Editors needed 46 seconds *less* than the Translators to complete the task, and a smaller difference of 9 seconds was noticeable for the excerpt from the novel, where the Post-Editors also needed less time than the Translators to complete the translation. It has to be emphasized that cumulative figures regarding task duration were not compared among the three text types, as they were not of exactly the same length.

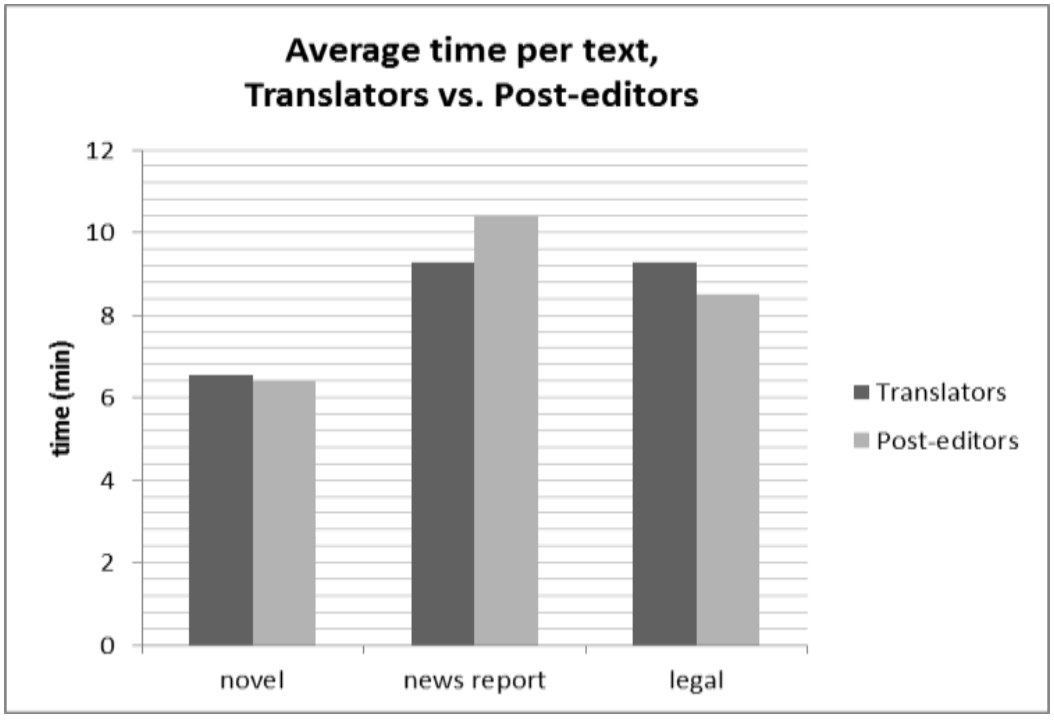

Figure 1

6.2 Quality of translations across texts

Although the time variable plays a significant role in the process of translating, and professional translators are almost without exception faced with tight deadlines, it is not only speed that counts. What is as important is the quality of translations. In this study, each translation was therefore compared in terms of the scores it received from the Evaluators, taking account of each parameter separately. Average scores were calculated for each translation, and for each of the four parameters.

6.2.1 Text 1 (novel)

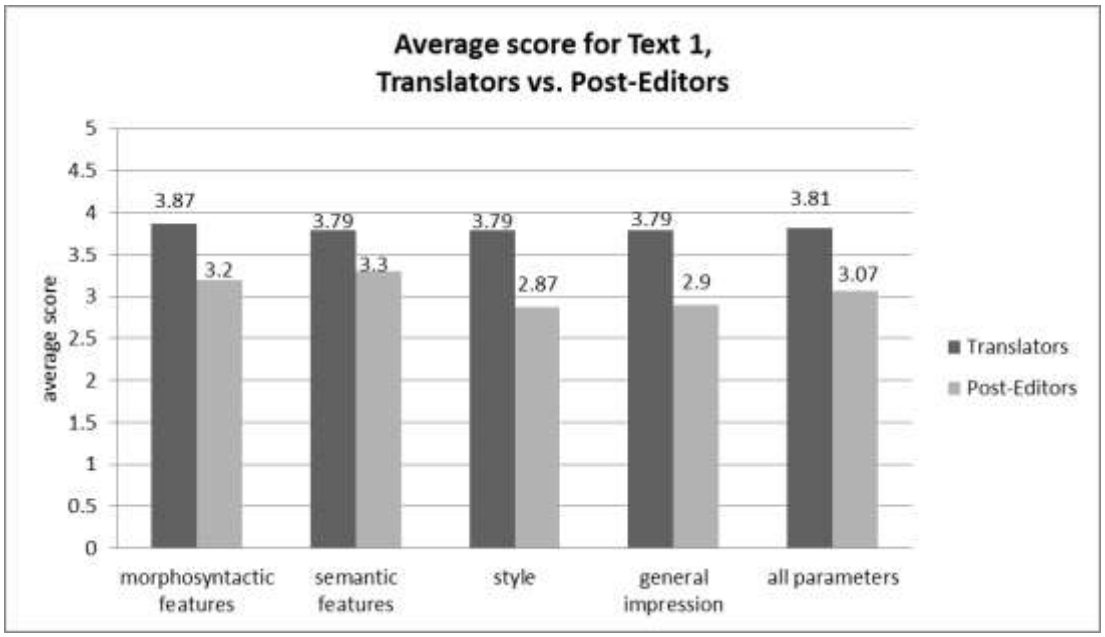

Figure 2

For Text 1, an excerpt from a novel, the Translators achieved better scores for each parameter (Fig. 2). With regard to the morphosyntactic parameter, on the scale of 1 to 5, the Translators' work was graded 3.87, while the Post-Editors' score was 3.2. With regard to the semantic parameter, the difference was smaller (0.49). The scores revealed that the biggest differences between the Translators and the Post-Editors referred to the stylistic parameter (0.92) and the general impression (0.89). These results showed that, even though the Translators did need some more time to perform their task (Fig. 1), the scores

they achieved justified that fact, especially when it came to style and the general impression. As the Translators needed on average only 9 seconds more than the Post-Editors to deliver their better-quality translations, it could be concluded, as expected, that Google Translate was not of much help when it came to translating texts belonging to fiction.

6.2.2 Text 2 (news report)

The results for Text 2, a news report excerpt, were quite different and rather unexpected (Fig. 3). For this text, the Post-Editors achieved better scores for each parameter, although the differences between them and the Translators were not as big as for Text 1. The differences with regard to the semantic parameter (only 0.1) and the general impression (0.23) can be considered negligible. There were bigger differences with regard to the morphosyntactic parameter (0.41) and style (0.43). However, we must not forget that the Post-Editors needed, on average, a whole minute and 8 seconds more to perform their task, which diminishes their success when it comes to the quality of the translations they delivered.

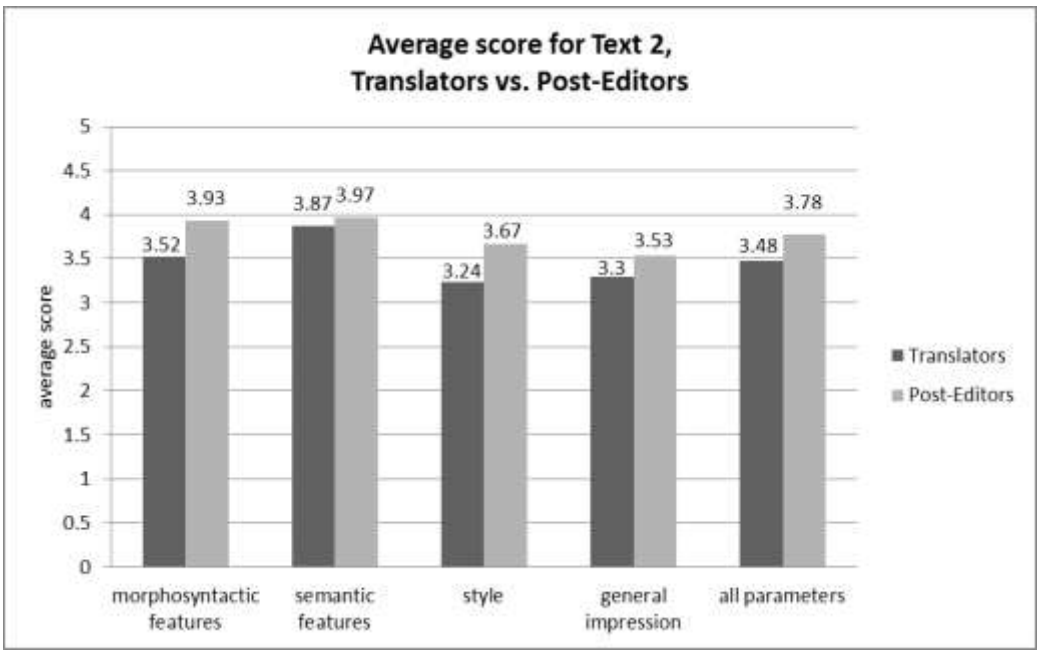

Figure 3

For news reports, therefore, Google Translate proved to be of some help; however, it seemed to slow down the process of translating, and the differences in quality were not that big. Nevertheless, the fact that the Post-Editors got better scores for this text for every parameter is quite interesting. This could be explained by the relevance of the topic. Since reports on the Fukushima tragedy were written, published, and translated worldwide, GT must have had a good selection of such texts in its parallel corpora, which made the job easier for the Post-Editors to a certain extent. With respect to both the time variable and translation quality, the question remains whether machine translation is more helpful or distracting when it comes to translating texts belonging to this genre.

6.2.3 Text 3 (legal)

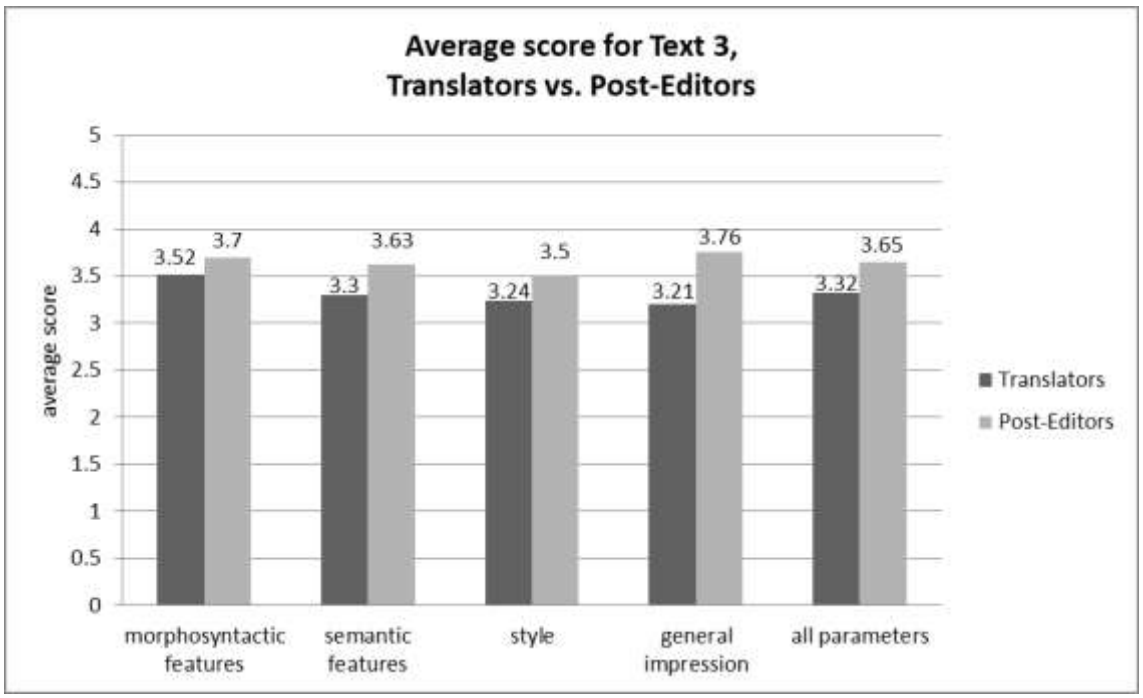

Figure 4

The results for Text 3, a legal text, confirmed my expectations: the Post-Editors achieved higher scores for each parameter (Fig. 4). However, the difference between their overall scores and those of the Translators was not as pronounced. The biggest difference was 0.55 for general impression. Smaller differences in achieved scores were found with regard to the semantic parameter (0.33), style (0.26), and the morphosyntactic parameter (0.18), the last being negligible.

Taking into account the fact that the Post-Editors performed their task considerably faster than the Translators, and having in mind that the Post-Editors delivered translations of better quality according to all four parameters, it could be concluded that, as expected, Google Translate did in fact prove to be of help in such translation tasks.

With regard to all three texts and all parameters, the biggest difference in scores had to do with style and the general impression for Text 1 (novel).

6.3 Quality of translations per parameter

In order to get a better insight into how the quality of translations varied according to given parameters, I also grouped results for every parameter across texts.

6.3.1 Morphosyntactic features

Figure 5 shows the results that reveal the quality of translations with respect to morphosyntactic features across texts.

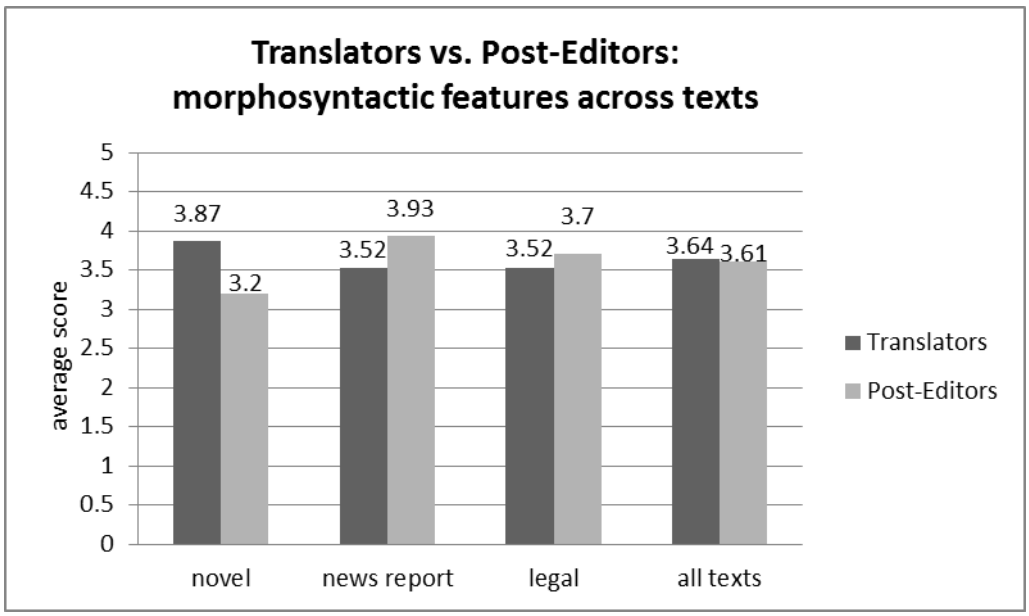

Figure 5

The comparison of scores across texts revealed that the most pronounced difference between the Translators and the Post-Editors on the morphosyntactic

level was evident in Text 1 (novel), and it amounted to 0.67 in favour of the Translators. It was followed by a smaller difference for Text 2 (news report), 0.41 in favour of the Post-Editors, and 0.18 for Text 3 (legal), also in favour of the Post-Editors. Bearing in mind the complexity of the Croatian morphology in comparison to the English morphology, my expectations for Google Translate's usefulness on the morphosyntactic level were not high, especially for Text 1 (novel), due to the specific nature of the language usually found in such texts. The results showed that such expectations were justified, since the Translators achieved a considerably higher score for Text 1 (novel).

As for Text 2 (news report), Google Translate did prove to be useful, as the Post-Editors achieved a higher score. This was somewhat surprising, since I expected the results on this level to be more in favour of the Translators, having in mind the nature of the language used in news reports, i.e. not as many fixed phrases, specific terminology and specific sentence structure as, for instance, in legal texts. However, this might be justified by the relevance of the topic of Text 2 (news report). As previously mentioned in section 6.2.2, the news on the Fukushima tragedy attracted global media coverage. Having in mind the way GT works, it might just be the case that in its database it had a lot of similar texts, and this fact resulted in a better quality of translation on the morphosyntactic level. The results for Text 3 (legal) were most surprising. Due to the specific nature of language usually occurring in legal texts, I believed it was justified to expect the results for Text 3 (legal) with regard to this parameter to be more in favour of the Post-Editors, as compared to the results for Text 2 (news report).

6.3.2 Semantic features

As for semantic features, the results shown in Figure 6 indicate that the highest difference (0.49 in favour of the Translators) between the Translators and the Post-Editors was again found in Text 1 (novel). This could also be justified by the way GT functions, as well as by the nature of the language usually used in novels – figures of speech such as metaphors and metonymies, comparisons, picturesque language, etc. might "confuse" GT to some extent, which might result in translations of poorer quality on the semantic level. The difference between the Translators' and the Post-Editors' scores for Text 2 (news report)

was negligible (0.1 in favour of the Post-Editors), which might point to the conclusion that the language in Text 2 was straightforward and clear, and that there were no problematic expressions that might have affected the quality of GT's translation on the semantic level. Again, the familiarity of the topic of the text might have also affected the results to be more in favour of the Post-Editors.

The Post-Editors achieved higher scores than the Translators for Text 3 (legal), which was expected, due to the fact that GT has access to a great array of parallel legal texts. The phrases and expressions used in such texts are almost invariably translated into the target language in the same way. However, the difference of 0.33 between the Post-Editors' and the Translators' scores was not that pronounced.

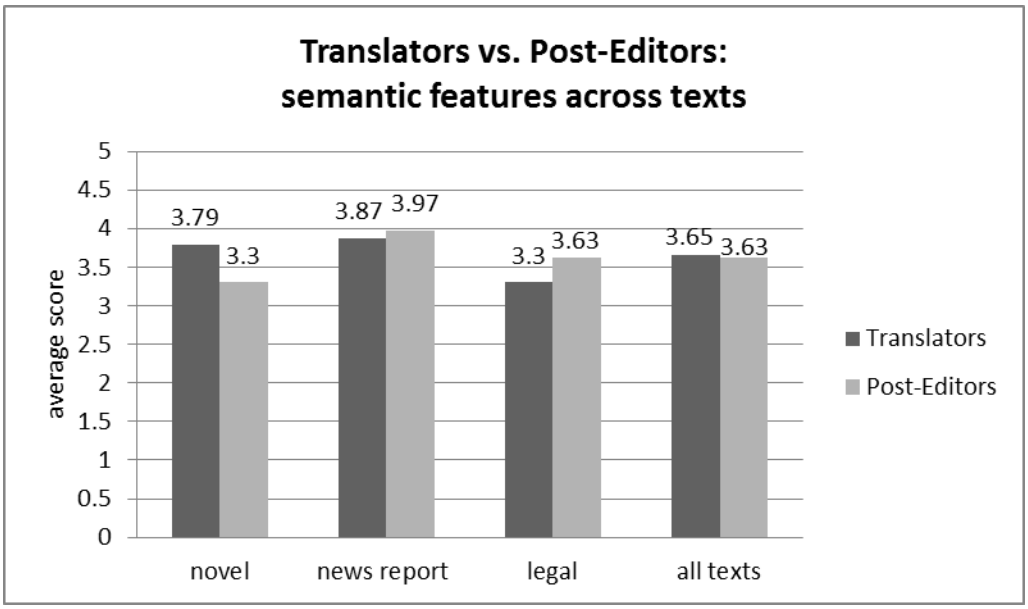

Figure 6

6.3.3 Style

When it comes to the stylistic parameter, there was a notable difference (0.92) between the Translators' and the Post-Editors' scores for Text 1 (novel) in favour of the Translators. This was expected, since style is a very important aspect of novels and other literary texts. Sometimes it is hard to capture and translate features of style even for human translators. For this reason, it was to be expected that the Post-Editors' scores for this parameter would be notably lower

than those of the Translators. The specific choice of words, sentence structure and their flow in this type of text is hard to recognize for a computer. In all likelihood, the translation produced by GT was particularly distracting for the Post-Editors and made this part of the task difficult for them.

The Post-Editors did, however, achieve higher scores than the Translators for both Text 2 (news report), 0.43, and Text 3 (legal), 0.26. Nevertheless, the differences were not as pronounced as for Text 1 (novel). The style of legal texts is also specific, but in a different way than fictional texts. It is more formal, and very strict rules are applied. Therefore, it was easier for GT to transfer the features of legal style into the TL than to do so when translating fiction.

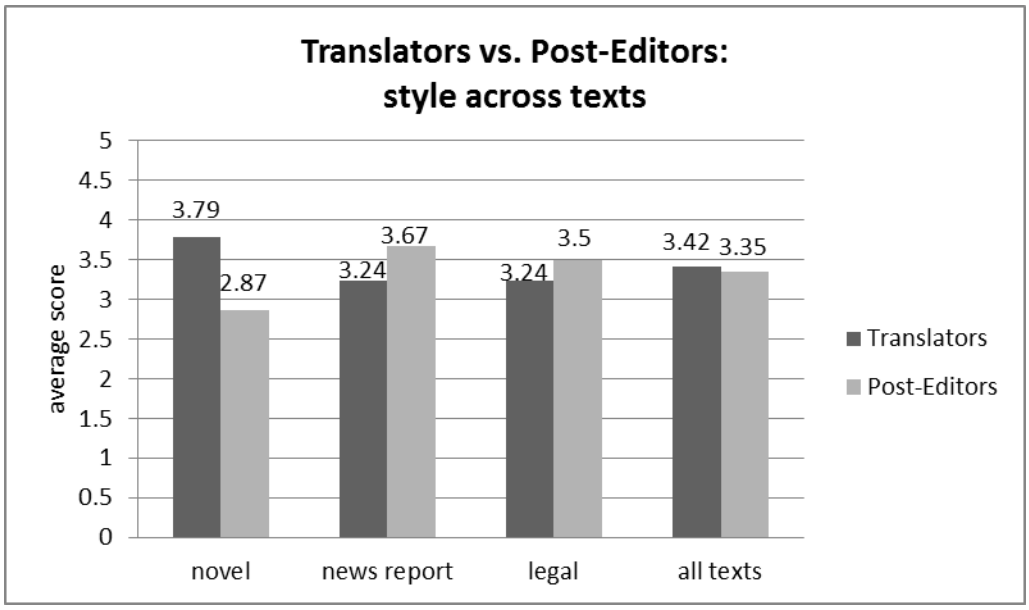

Figure 7

6.3.4 General impression

The results referring to the general impression parameter were quite similar to those referring to style. The Translators were more successful in Text 1 (novel) – there was a notable difference in scores of 0.89 – while the Post-Editors were more successful in the other two texts. For Text 2 (news report), the difference in scores was not as notable (0.23), and there was a bigger difference (0.55) for Text 3 (legal). The explanation for such results is similar to that referring to the style parameter.

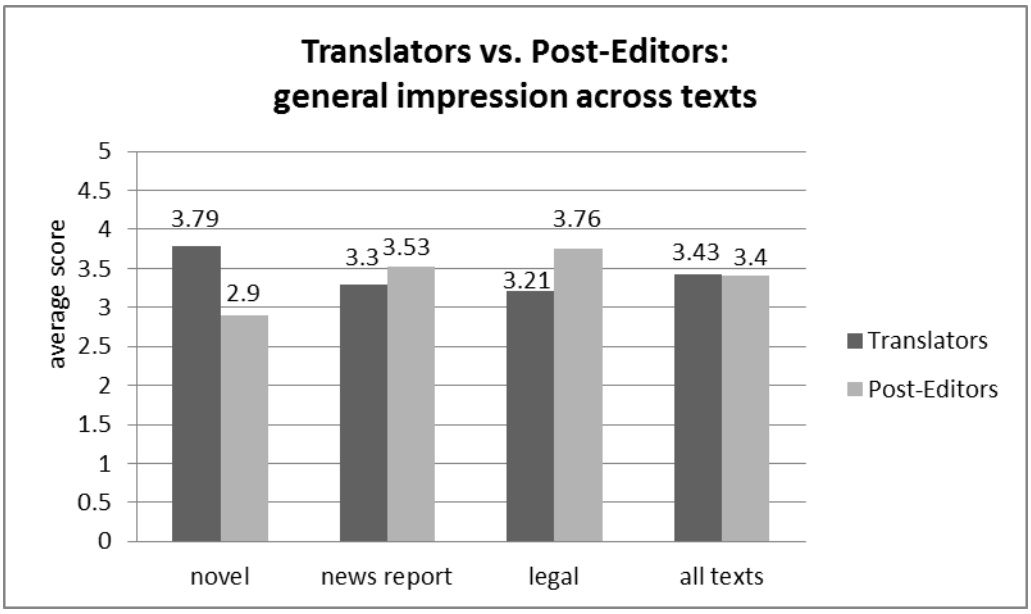

Figure 8

7. Discussion

The results referring to the time variable showed that the Translators completed their task faster than the Post-Editors, which refuted my first hypothesis. Such findings must have resulted from the fact that the participants in the experiments did not use computers. As it was, the Post-Editors needed more time to complete their tasks than they would have if they had been able to edit the MT output on screen, without having to rewrite the translations.

Furthermore, although the difference of 17 seconds between the two groups seemed negligible, a more detailed insight into the results for each text with respect to the time variable revealed that the difference between the two groups in fact did vary from one text type to another. My expectations were confirmed in the case of Text 1 (novel) and Text 3 (legal), with the differences between the Translators and the Post-Editors of 9 seconds and 46 seconds, respectively, the Post-Editors being faster. In contrast, the results referring to the time variable for Text 2 (news report) were different from what I had expected, since the difference between the two groups was 1 minute and 8 seconds, with the Translators being faster. This might have been the result of the confusing and

misleading nature of machine translation output, given that the Post-Editors spent much more time correcting the errors from the MT output than the Translators needed to produce the whole translation from the very beginning.

A more careful look at the results concerning the quality of translations indicated that the Translators obtained, as expected, better scores for each parameter for Text 1 (novel), while the Post-Editors produced translations of better quality with respect to all four parameters for both Text 2 (news report) and Text 3 (legal).

Taking both the time variable and the quality of delivered translations into account, machine translation did not prove to be of much help for the translation of the excerpt from a novel. Although slightly slower (9 seconds), the Translators produced translations of notably higher quality for every parameter in this text. For this reason, in the case of Text 1, the Post-Editors' advantage related to the time variable can be neglected. The results for Text 2 (news report) showed a much greater advantage for the Translators when it came to the time variable; however, it was the quality of translation that suffered. It was particularly true of the morphosyntactic features and style, and less so for the semantic features and the general impression. Given that, as previously mentioned, it is not only the speed that counts, but the quality of translations as well, it could be concluded that machine translation was helpful for this type of texts. In the case of Text 3 (legal), my expectations were completely confirmed: not only were the Post-Editors 46 seconds faster than the Translators, but they also produced translations of better quality for each of the four parameters.

8. Conclusions

The aim of this paper was to examine the extent to which MT systems are helpful in everyday translation tasks. The study set out to test the hypothesis that post-editing of MT output would be faster than fully human translation, but of inferior quality. To be more precise, I hypothesized that post-editing of machine translation output would prove more helpful for the translation of texts tending to use predictable language than texts that use language in a more creative way. In line with this expectation, post-editing of machine translation output was

found to be faster than fully human translation in the case of texts belonging to fiction and legal texts; however, human translation of the news report was in fact found to be faster than post-editing of MT output of the same text.

Regarding the quality of translation, my hypothesis was confirmed in that the results of my research showed that texts tending to use language more creatively (in this case, an excerpt from a novel) do not seem to benefit from the use of MT tools. On the other hand, when it comes to texts belonging to more restricted semantic domains, where predictable language is used, translators can benefit from the use of MT tools, since post-editing of such text was found in my research to require less time than fully human translation, and the quality of the final product was higher than that of fully human translation. In this study, the news report partly confirmed, and partly refuted my hypotheses. Fully human translation required less time for completing the task than post-editing of MT output, but the quality of post-edited MT output was higher. As mentioned in previous sections, the explanation for such findings could be found in the fact that the topic of the article used in the study had been very well covered in the international media, which in turn resulted in statistical MT output of better quality in the first place. Further research might explore this issue in greater detail. For instance, it would be interesting to conduct similar research comparing fully human translation with post-editing of MT output in terms of speed and translation quality, using two different news reports: one reporting on such a global event covered by the international media, and the other, perhaps from a local newspaper, reporting on a minor, very culture-specific event, for which it would be unlikely to find similar texts in MT parallel corpora.

8.1 Limitations of the study and future research

Although I planned each stage of the research thoroughly, I did come across some limitations that might have influenced the results, and, indirectly, the conclusions offered in this paper.

I find it important to stress that the conclusions above are valid in the context of this particular study. Having in mind the limitations of the research, I am aware that they might not have broader ecological validity. However, the fact

that some of the findings chime with those of other researchers dealing with the same matter suggests that I was going in the right direction. The main limitation of the study has to do with the fact that the participants did not work on computers, and this probably influenced the results, especially for the group of participants who were asked to revise GT translations. Had they worked on computers, I presume this group of participants would have had better results referring to speed, because they would have been able to simply correct GT mistakes on the screen, and they would not have had to rewrite the translations from the beginning. Also, as previously mentioned, GT now has a feature that allows users to click a word or word phrase in GT translation and view other possible translations, which was, naturally, not the option in this research, and it would have accelerated the process of post-editing to a certain extent. Thus, I believe this limitation influenced the results the most.

Another limitation that I find important to mention is that the participants in this research were not professional translators with a lot of practice in such assignments, which also might have affected the results. However, having in mind that neither of the two groups of participants were professionals, I believe that this limitation did not influence the overall results to such an extent as the first one. Furthermore, the participants were asked to make a note of the time they started and finished working on each translation themselves, which might have opened the possibility of manipulating the results. In addition, the number of participants was relatively small, and in order to obtain more accurate results, a similar study on a bigger sample should be conducted.

In addition, the texts used for this research were quite short, and it would be interesting to compare the results of this study to those of a similar one using longer texts. Finally, it is important to mention the evaluators' relative objectivity when grading translations. This is why they were asked to grade all translations in one sitting, with short pauses in between, in order to reduce the effects of external factors such as fatigue, current mood, etc. Furthermore, the evaluators who participated in the research had not had extensive professional experience. In order to accomplish a higher level of objectivity, a similar study in which more experienced translators would participate should be conducted.

In conclusion, the results of this research show that MT tools, or rather, Google Translate, can be helpful when it comes to translation of certain types of texts. However, it is very important that post-editors know both the source and target languages very well, as solutions offered by such systems might be misleading or completely wrong, or at least stylistically inappropriate. Moreover, such systems could also be used by translators to obtain ideas for possible translations and not to simply copy-paste their solutions. Possibly, in the future we will see a technological leap enabling fully automated machine translation applicable to a wide range of texts, without the necessity for post-editing such translations by humans. However, for the time being, there is still a long way to go until this goal is reached. Despite the existence of fully automated machine translation systems that function well for texts using more restricted semantic domains and controlled language, interventions of the human mind in various aspects of the translation process are still a necessity.

References

Bellos, David. 2011. *Is That a Fish in Your Ear? Translation and the Meaning of Everything.* New York: Faber and Faber.

Calude, Andrea. 2014. "Machine Translation of Various Text Genres." 7th Language and Society Conference of the New Zealand Linguistic Society, November 2002, Hamilton, New Zealand. URL: http://www.calude.net/andreea/MT.pdf. Accessed on 15 May 2014.

Council of the European Union. 2008. "Council Resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)". *Official Journal of the European Union C 320/01*. URL: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008G1216(01). Accessed on 12 May 2012.

Douglas, Arnold, Balkan, Lorna, Meijer, Siety, Humphreys R. L. and Sadler, Louisa. 1994. *Machine Translation: An Introductory Guide*. London: NCC Blackwell. URL: http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/. Accessed on 7 April 2012.

European Comission. 2014. "Translation and the European Union." URL: http://ec.europa.eu/dgs/translation/translating/index_en.htm. Accessed on 20 June 2014

Fiederer, Rebecca and O'Brien, Sharon. 2009. "Quality and Machine Translation: A realistic objective?" *The Journal of Specialised Translation* 11. URL: http://www.jostrans.org/issue11/art_fiederer_obrien.pdf. Accessed on 22 May 2014.

Google. 2014. "Google Translate." URL: http://translate.google.com/about/. Accessed on 26 May 2014.

Hutchins, John. 1995. "Machine Translation: A Brief History." In: Koerner, E.F.K. and Asher, R.E. (eds.), *Concise history of the language sciences: from the Sumerians to the cognitivists.* Oxford: Pergamon Press. 431-445. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.7297&rep=rep1&type=pdf. Accessed on 10 Feb. 2014.

Hutchins, John. 1997. "Fifty years of the computer and translation." *MT News International 16:* 14-15. URL: http://hutchinsweb.me.uk/MTNI-16-1997.pdf. Accessed on 14 Feb. 2014

Hutchins, John. 2012. "The development and use of machine translation systems and computer-based translation tools." International Symposium on Machine Translation and Computer Language Information Processing, 26-28 June 1999, Beijing, China. URL: http://hutchinsweb.me.uk/Beijing-1999.pdf. Accessed on 10 Apr. 2012.

Koponen, Maarit. 2010. "Assessing Machine Translation Quality with Error Analysis." In: *Electronic proceedings of the KäTu symposium on translation and interpreting studies 4*. URL: http://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf. Accessed on 24 April 2014.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik.* Zagreb: Ex Libris.

Uszkoreit, Jakob, Ponte, Jay M., Popat A. C. and Dubineret, Moshe. 2010. "Large Scale Parallel Document Mining for Machine Translation". *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 1101–1109. URL: http://aclweb.org/anthology//C/C10/C10-1124.pdf. Accessed on May 17 2012.