Comparing Post-Editing Difficulty of Different Machine Translation Errors in Spanish and German Translations from English

Anna Zaretskaya[a]
Mihaela Vela[b]
Gloria Corpas Pastor[a]
Miriam Seghiri[a]

a) University of Malaga
Campus de Teatinos
Facultad de Filosofía y letras
Bulevar Louis Pasteur s/n
{annazar, gcorpas, seghiri}@uma.es

b) Saarland University
Campus A2 2, Room 1.14 D-66123
Saarbrücken, Germany
m.vela@mx.uni-saarland.de

Abstract

Post-editing (PE) of Machine Translation (MT) is an increasingly popular way to integrate MT in the professional translation workflow, as it increases productivity and income. However, the quality of MT is not always good enough to blindly choose PE over translation from scratch. This article studies the PE of different error types and compares indicators of PE difficulty in English-to-Spanish and English-to-German translations. The results show that the indicators in question 1) do not correlate between each other for all error types, and 2) differ between languages.

Key Words: post-editing, Machine Translation, MT errors, CAT tools.

1. Introduction

The developments in the area of Machine Translation (MT) have drawn the attention from the translation industry, which since then has been in continuous search for the optimal way to incorporate MT in the professional translation workflow in order to increase incomes. Post-editing (PE) of MT is one of the ways to use MT that is currently being adopted by more and more translation service providers (TSPs). PE is usually understood as "a human being (normally a translator) comparing a source text with the machine translation and making changes to it to make it acceptable for its intended purpose" (Koby, 2001, p. 1). As TSPs turn to PE, it also becomes a topic of growing interest in in the academia. Some research works have proven PE to be, in many cases, more effective than translating from scratch and to boost translators' productivity (Laübli et al., 2013; Zampieri & Vela, 2014; Zhechev, 2014). It is still recognised, however, that PE is not suitable for any machine-translated text and for any translation scenario. Depending on the purpose of translation, language combination, text domain, genre, and MT engine, the MT output can sometimes turn out to be more difficult to post-edit than to translate the source sentence from scratch. Apart from that there is a variety of different types of MT errors, and we suggest that some of them are easier to post-edit than others.

This article aims at investigating how different MT errors influence the post-editing process in two target languages. More specifically, it analyses results of a post-editing experiment, which consisted of post-editing sessions of English-to-German and English- to-Spanish translations. The post-edited segments were previously annotated for errors using a specific error taxonomy. We compare the two languages as to different indicators of

PE difficulty obtained for each error type. In other words, we want to find out whether the same errors are difficult or easy to post-edit in both languages. According to Krings (2001), there are three types of post-editing effort: temporal, cognitive, and technical. In our research, we only consider the temporal and the technical types. Temporal effort refers to the time one person takes to post-edit a given segment, which we will further refer to as post-editing time, or PE time. Technical effort consists of the changes made in the MT output to obtain the final version. It can be represented by the number of keystrokes or the number of deletions, insertions, and substitutions made. We will use the post-editing effort (PEE) measure, which is a numerical indicator of the amount of editing made within the segment, and is based on fuzzy match algorithms used in CAT tools. In other words, PEE roughly shows what part of the segment was edited in relation to the whole segment. In order to avoid confusion, we will further use the term *post-editing effort (PEE)* to refer to this specific measure of technical post-editing effort as opposed to the broad sense defined by Krings (2001), which we will further refer to as *post-editing difficulty*.

The remainder of the article is structured as follows. Section 2 presents relevant work on the subject of MT errors and post-editing, paying special attention to comparison between languages. In Section 3 we present our methodology, including the experiment setup (3.1) and the data we used (3.2), followed by the results presented in Section 4. Section 5 will conclude the findings and outline some ideas for future research.

2. Related Work

A big part of PE research is focused on investigating whether it is a beneficial practice and the best translation scenarios to use it (Federico et al., 2012; Parra Escartin & Arcedillo, 2015). Even though considerable research has been carried out on the post-editing process in general, few works take into account the language differences, i.e. that the PE process can be different in different languages. Indeed, it is believed that MT, and consequently PE of MT, is more useful for some languages than other, mostly because of the quality of MT. The reason why such comparison has not been performed is mainly that it is rather difficult to find comparable data for post-editing experiments in different languages.

To our knowledge, only Popovic et al. (2014) considered multiple language pairs in their PE experiments. They explored the relations of five different types of edit operations with the cognitive and the temporal PE effort. Cognitive effort in this case was represented by manually assigned difficulty scores. The operation types were based on edit distance and included correcting word form, correcting word order, adding omission, deleting addition, and correcting lexical choice. The data consisted of French to English, English to Spanish, German to English and English to German translations. The study included the analysis of 'almost acceptable' translations, i.e. translations that require only little editing. Those segments were filtered out and the number of different kinds of edit operations in the language pairs in question was compared.

The authors found that the lexical edits (such as correction of mistranslations) were the most prominent for all translation directions. Word Form errors were rare in English out- puts, but still relatively high in Spanish and German translations. As for word order errors, for French-English and English-Spanish translations the reordering edit rates were low, however for German-to-English translations they were much higher. This high rate indicates that, for this translation direction, even high-quality translations contain a significant number of syntactic errors. Apart from that, similarly to the research reported in (Koponen, 2012), word order errors correlated most strongly with the perceived cognitive effort, together with mistranslations, and that the lexical errors correlated mostly with post-editing time.

There has been done other research on different error types in post-editing, however without consideration of different languages. These works mostly involve measuring different PE effort indicators for different error types. Namely, Temnikova (2010) proposes a ranking of MT errors according to their cognitive difficulty, or, in other words, to how cognitively difficult she considers them to be for post-editing. Following the same line, Koponen et al. (2012) also provide a ranking of MT errors. The authors base their ranking on their own judgment. Lacruz et al. (2014) also look at different types of MT errors, but from the point of view of the concepts of cognitive demand and cognitive effort. Cognitive demand is closely related with MT utility and expresses the usefulness of MT output for producing correct translation. Cognitive effort is the actual effort exerted during the post-editing and is similar to the technical post-editing effort. They found out that mistranslations, omissions, additions and syntax errors have stronger correlation with cognitive demand and cognitive effort (for instance, as indicated by pause to word ratio) than "less cognitively challenging errors" such as punctuation and word form errors.

Daems et al. (2015) introduce into their measurements more different indicators of PE difficulty, in particular average number of production units, average duration per word, average fixation duration, average number of fixations, pause ratio, average pause ratio. They used linear mixed models in order to calculate the effect coefficient. In this study, both the classification of errors and the annotation was performed by the authors.

One of the main findings of this study was that the most common error types affecting PE difficulty indicators are mistranslations, structural issues, and word order issues. But more importantly, different error types affect different PE effort indicators. Thus, more technical errors such as mistranslation, grammar, structure, word order affect the pause ratio, the average pause ratio, and the average number of production units, which are therefore related between each other. The average duration per word is affected most by coherence and structure issues. Fixation duration is strongly related to mistranslation errors. The average number of fixations is predicted by coherence issues, which is a more cognitively demanding error type. It is not always clear, however, how the effects of different error types intervene with each other, i.e. when there are several error types in one segment we cannot know in which exact place the fixations or the pauses took place.

3. Method and data

For our experiments, we used corpora of English sentences translated into German and Spanish with different MT engines and subsequently annotated for translation errors (Burchardt et al., 2013). We chose only sentences that contained one error, which allowed us to separate the effects of each error type on the post-editing process. The location of the errors was indicated using braces to ensure that the students post-edited only the same strings that were annotated as erroneous, as shown in the example below. In other words, we wanted the students to correct only the annotated errors.

(1) Active Orthodox Churches
Iglesias ortodoxas {activo}

Identification of the MT errors is an important step in the post-editing process. As demonstrated by Valotkaite and Asadullah (2012), prior highlighting of errors for post-editing make editors more efficient as they miss less error. Nevertheless, we had to discard error identification step in our experiment in order to limit the editing area of the segment to be corrected. This way, our experiment is more controlled, but on the other hand, it is more distant from a natural post-editing setting that professional translators work in. Another difference between our experiment and the common professional post-editing setting was that segments were not related to each other, which made it more difficult for the students. In fact, in the informal feedback we got from them, they commented that some segments were hard to translate without knowing the context. This limitation is due to the corpus, as well as the condition of having only one error per segment, which is impossible to have in an entire text.

3.1 Experiment setup

The experiment participants were 19 native German speakers enrolled in the translation programme, 12 of which were doing their last year of bachelor course, and 7 were master students, and 24 native speakers of Spanish, of which 6 were doing a master course, and 18 were in their last year of bachelor degree. Before starting the experiments, the students were given written instructions explaining that they should correct, if possible, only the segment parts between the braces. They were also instructed to make only the corrections necessary to achieve a grammatically and semantically correct translation. In addition, they completed a prior short exercise, during which they could practice and prepare for the task.

The CAT tool used for the experiment was Matecat.[1] It was given preference for the editing log feature it provides, which allows to see all the corrections made and registers different statistical information about the translation process, including the PE time and PEE for each segment. In addition, Matecat is web-based, which made it easy for the students to access their translation jobs without any prior installation, and at the same time it gave us control over the post-editing process. Finally, it is free of charge and very easy to use. In order to distribute the sentences among the participants we, first of all, randomly selected ten sentences in German and ten sentences in Spanish that were to be included in every student's translation job.

This was done in order to measure the annotators' agreement. The rest of the selected sentences from the corpora were divided into sets, each set had 48 or 49 segments in German and 50 or 51 segments in Spanish.

[1]https://www.matecat.com/.

This way, each sentence except for the ten common sentences was post-edited by four or five students in German and by seven to nine students in Spanish. Students received a URL with their personal Matecat job which included a document to translate and a TM containing the annotated translations. All translation units in the TM had 100% match score.

In this kind of task, it is natural that the participants take more time to edit the first segments compared to the last ones, when they are already more familiar with the process. As editing time is an important variable in this research, we tried to avoid this limitation by randomising the order in which the segments appear. In addition, the students were given a test exercise with 5 sentences to get accustomed with the task before starting the actual experiment.

3.2 Corpora

The data used for the experiments was selected from the MQM error annotation corpora (Burchardt et al., 2013). The MQM corpora contain English-to-German and English-to-Spanish translations produced by statistical, rule-based and hybrid engines. The corpora were designed to contain sentences that exhibited only few errors, or almost acceptable translations. The English source sentences are different in the German and Spanish corpora; however, they come from the same data sets. One part of the corpora originated from previously existing publicly available corpora of machine-translated texts from technical domain and news (Avramidis et al., 2012; Specia, 2011). The second part was taken from the TSNLP Grammar Test Suite[2] for English. All the data are publicly available online in XML format.[3]

The translations in the corpora were annotated for errors by up to five language professionals. The annotation was performed according to the Multidimensional Quality Metric (MQM), developed within the QTLaunchPad project.[4] The metric provides a method for translation error annotation for various purposes and with various degrees of granularity (Lommel, 2013), and contains an error taxonomy as well as guidelines for annotation.

For investigating usability of different types of MT errors for post-editing, almost acceptable translations are of special interest. Firstly, because, as pointed out above, very bad translations are not worth editing. Secondly, bad translations tend to have various errors which are often hard to separate from each other. For this reason the data chosen for the experiment consisted exclusively of sentences with only one error. In sentences where the error types were different among annotators, we chose the more frequent version. This way we selected 200 sentences from the English-to-German corpora and 163 sentences from the English-to-Spanish corpora. The total number of source words in the German corpus amounted to 1941, which makes it about 10 words per sentence on average. The Spanish corpus contained 2347 source words, which makes it 14.4 words per sentence on average. The selected sentences were then extracted from the XML files.

We used a version of the error taxonomy that was less granular, in order to avoid having error types with very few examples in the corpora. The taxonomy was the following:

- Accuracy
  - Mistranslation
    - Terminology
    - Overly Literal
    - Number
    - False Friend
    - Entity
    - Should Not Have Been Translated
  - Locale Convention
  - Inconsistency
  - Omission
  - Addition
  - Untranslated
- Fluency

[2]http://www.delph-in.net/tsnlp/ftp/tsdb/ •.
[3]http://www.qt21.eu/deliverables/annotations/index.html, http:// www.qt21.eu/deliverables/test- suite/.
[4]http://www.qt21.eu/launchpad/.

- Style/Register
- Unidiomatic
- Spelling
- Typography
- Grammar
    - Word Form
    - Word Order
    - Function Words
- Unintelligible

Not all the error types were present in both corpora. Consider Table 1, which shows the error distribution for each language pair.

4. Results

4.1 General statistics

We obtained 19 translation sessions for the English-German language pair and 24 sessions for English-Spanish language pair. The average PE time for German was $time_{avgDE}$= 40.88 seconds, while in Spanish it was a little higher: $time_{avgES}$= 54.35. We assume that the reason is the difference in the average sentence length mentioned above. The average PEE in German was equal to PEE_{avgDE}= 0.24 while for Spanish it turned out to be only PEE_{avgES}= 0.12.

We calculated the inter-annotator agreement in order to find out how similar the students were in terms of PE time and PEE (Table 2). It was calculated based on the 10 common segments post-edited by all the participants in each language combination. The measure used was the Intraclass Correlation Coefficient (ICC) (Shrout & Fleiss, 1979), which takes values between -1 and 1. The agreement was significantly higher for Spanish both in PE time and PEE. We obtained the ICC values for time ICC_{timeES} = 0.38 and ICC_{timeDE} = 0.16; and for post-editing effort ICC_{PEE-ES} = 0.53 and ICC_{PEE-DE} = 0.36. The reason for this is probably that the Spanish students followed the instructions more carefully, but also that the Spanish experiment took place after the German, and that is why it was better planned and controlled. In order to see how the two difficulty indicators are related, we calculated the correlation between the post-editing time and PEE using Pearson's correlation coefficient, and we obtained generally weak correlation, although it was slightly higher for the Spanish data: r_{es}=0.15, r_{de}=0.07.

4.2 Comparing different error types

We compared the average time and PEE for each error type obtained for both languages. It has to be mentioned that, even though the error classifications were very similar in the two experiments, there were still some differences in error types that were present in the two corpora, which we saw in Section 3.2. For this analysis we selected only the errors present in both German and Spanish corpora. First of all, we discovered that edit time was higher for Spanish and PEE was higher for German across most of the error types. Further analysis revealed that both time and PEE strongly depend on the target segment length. In fact, previous studies already demonstrated the relation between PE time and segment length (Popovic et al., 2014; Koponen, 2012). We obtained Pearson's correlation coefficient between sentence length and average time equal to r=0.47 for Spanish and r=0.26 for German. We also observed a strong negative correlation between segment length and average PEE: r_{es}=-0.46 and r_{de}=-0.37. This means that in many cases, longer segments take more time to edit, and tend to have smaller PEE. Note, that PEE is a measure that directly depends on the segment length, as it roughly represents the percentage of the part of the segment that was edited, compared to the whole segment in number of words. That is why we obtained negative correlation between PEE and the segment length: the longer the segment is, the smaller is the part covered by the edited chunk. Indeed, the segments in the Spanish corpus were significantly longer.

4.2.1 Post-editing time

Keeping this in mind, we analyse the differences between the two languages in the average post-editing time calculated for each error type (Figure 1). Closer analysis of the data confirmed that the reason for these differences was mainly the sentence length. Thus, the biggest difference in the average edit time was observed in the unintelligible error type: the time in Spanish was significantly longer. At the same time, the Spanish segments were also much longer than the German ones, in particular, the average length in Spanish for this error type was 30 words and in German only 12 words per segment.

Similar differences in segment length are observed in all error types where Spanish students took significantly longer time, such as Word Order, Untranslated and Grammar. The only two errors that took longer to post-edit in German were Mistranslation and Style/Register. The latter type was represented only by one sentence in the German corpus, which was longer than the Spanish sentences from the same category. Similarly, Mistranslation had an average length of 10.6 words in German, while the average length in Spanish was 16. In comparison, the Spelling errors showed very similar post-editing time, and the average segment lengths in this error were also similar: 9.6 in German and 9 in Spanish.

These numbers show which errors took specifically long time to correct in Spanish compared to German. This was the case for almost all errors, but especially for Word Order, Unintelligible and Untranslated. However, this does not mean that English-to-Spanish translations are harder to post-edit than English-to-German. In fact, sentence length was the main reason for the different results in Spanish compared to German. This is because, even though the erroneous part only covers a certain part of the sentence, a post-editor needs to read and understand the whole sentence, which can take different time depending on the length. Knowing that the sentences in the Spanish corpus were generally longer, we want to avoid this bias and use a measure that is independent of length in order to be able to compare the two languages. We suggest the time-per-word measure, which is calculated by the following formula, where *t* is the total time to edit a given segment, and *l* is the segment length in words:

$$t_w = t/l.$$

Average time-per-word is then calculated for each error type (Figure 2). This graph is very different from Figure 1. First of all, note that the Unintelligible errors are very similar in the two languages. On the other hand, big difference is observed in Mistranslation and Typography. Secondly, there is no consistency in which language has higher time-per-word values; it mostly depends on the error types. Thus, the errors where time-per-word is higher in Spanish are: Function words, Word Form, Word Order, Terminology, Style/Register, and Spelling. The errors where German took longer time are: Mistranslation, Omission, Unintelligible, Typography, Addition, Untranslated, and Grammar. It is interesting that the Word Order errors turned out to have higher time-per-word values for Spanish, even though Spanish has a more similar word order to English than German does.

4.2.2 Post-editing effort

The PEE obtained for the German corpus was generally higher than for the Spanish corpus (Figure 3). The only exceptions were the Word Order, Addition and Spelling categories. While the average length of the segments with Word Order errors was very similar in Spanish and German (15 and 13.75 respectively), there is a bigger difference in Addition: 15 and 6.7. Despite our assumption that longer segments are related to lower PEE scores, Spanish still shows higher PEE. The analysis of the editing logs in Matecat showed that this was due to an error in the program: some of the Addition errors were not retrieved from the TM; therefore, students used another TM suggestion automatically shown by the program, which had a lower fuzziness score. Thus, they had to correct significantly more text.

As to the Spelling errors, consider the following example.

(2) Next to the name, and choose a new name from the picker.
{Neben} dem Namen und wählen Sie einen neuen Namen aus der Auswahl.

It is a capitalization type of spelling error. This and the other two segments that belonged to the catipalisation type received lower PEE score than other spelling errors in German. The analysis of the correction operations made by the students in these segments showed that, in this case of capitalisation error, the students only corrected the misspelled letter. In other sentences with spelling errors, they mostly substituted the whole word. Considering this, we can make a conclusion that spelling errors related to capitalisation take less PE effort compared to other spelling errors. Major differences between the PEE in two languages were observed in Typography, Terminology, Omission, Untranslated and Grammar errors. Again, we assume that, like in the case of PE time, the differences might come from the difference in segment length. Typography errors were an exception, as the average PEE result was significantly biased by one of the PEE values, where the student changed the entire sentence (Figure 3).

The experiment results revealed some interesting findings about the relation between PE time and PEE. We already mentioned above that there was found only weak correlation between these two variables, with the Pearson's r values equal to 0.15 for Spanish and 0.07 for German. We have noticed, however, that the PEE graph looks much more similar to the graph of time-per-word values.

Indeed, the calculation of the Pearson's correlation coefficient proved our hypothesis: it increases from 0.15 to 0.69 in Spanish and from 0.07 to 0.33 in German with the time-per-word measure.

5. Conclusions and future work

In this article, we presented results of an experiment that aimed at comparing the difficulty of post-editing of different MT errors in English-to-German and English-to-Spanish translations. As indicators of post-editing difficulty we used PE time (temporal post-editing effort) and PEE (technical post-editing effort), as calculated by Matecat editing log. The sentences in the corpora of MT translations were previously annotated for MT errors, and the location of the errors was marked for post-editors, so that they did not need to look for the errors. Each sentence contained only one error.

We discovered that the PE time was on average higher for Spanish, while the PEE was lower. We came to the conclusion that this was due to the sentence length, which was generally bigger in the Spanish translations. Both of the measures we used, PE time and PEE, highly depend on the sentence length, which significantly influenced the results. Therefore, we suggested the average time-per-word measure, which reflects the average time taken to post-edit one word. Furthermore, for future similar experiments it is desirable to take this into account by selecting a corpus with similar sentence length.

We also investigated how the two indicators of post-editing difficulty are related. The question is whether we can actually talk about general difficulty of a certain type of errors, or we should rather talk about temporal and technical PE effort as two independent aspects of post-editing difficulty. There was only weak correlation between PE time and PEE in our experiment, while it was slightly stronger for Spanish than for German. On the other hand, time-per-word has a stronger correlation with PEE in both languages, and even very strong in Spanish. Therefore, we can conclude that time-per-word and PEE is much more related than PE time and PEE.

We compared the PE difficulty indicators for the types of MT errors that were present in both corpora. We discovered that the difficulty of errors varies significantly between the two languages, and also between the two difficulty indicators. For instance, errors that took long time to correct do not necessarily show high PEE scores. In addition, errors that took long time in one language could turn out to be fastest to correct in the other. Thus, based on the time-per-word measure, Mistranslation and Typography errors seem to take much more time in German than in Spanish. As to PEE, apart from these two errors, the biggest difference was also observed in Terminology, Untranslated, and Grammar, where German also showed higher difficulty scores. On the other hand, there are cases like Function Words, where we can observe higher time-per-word in Spanish, but lower PEE, or Addition, with higher time-per-word but lower PEE in German.

To summarise, only from comparing these two languages, we can see that the post- editing difficulty of different MT errors vary significantly across languages. Future re- search in this direction would involve carrying out a broader comparison with more languages, preferably with more structural and grammatical differences.

Figures

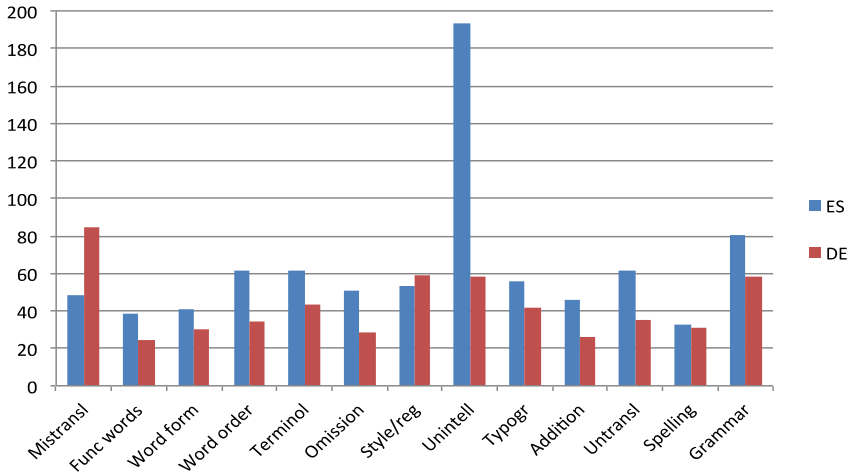

Fig. 1: Average post-editing time in Spanish and German by error type.

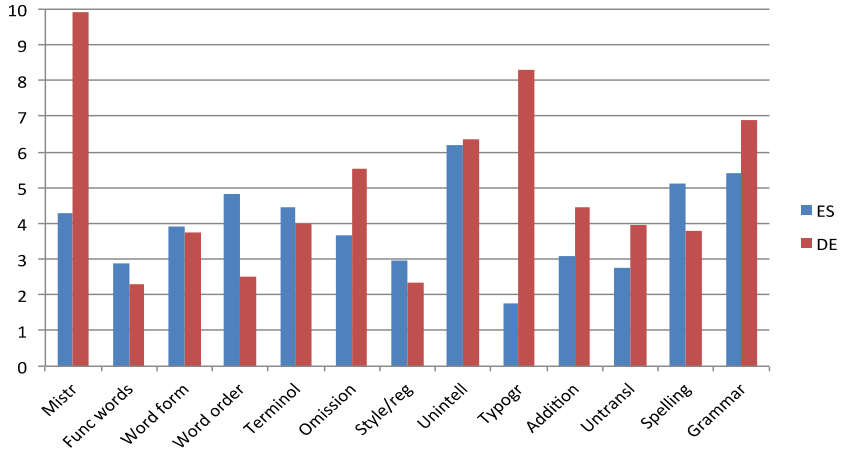

Fig. 2: Time-per-word for Spanish and German.

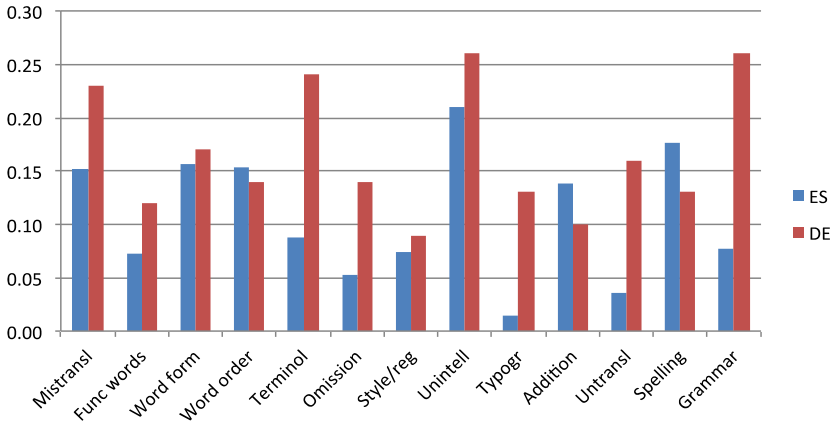

Fig. 3: Average post-editing effort in Spanish and German by error type.

Tables

Error type	Frequency DE	Frequency ES
Mistranslation	41	36
Word Form	34	24
Function Words	30	33
Overly Literal	18	0
Locale Convention	14	0
Omission	12	6
Spelling	9	2
Word Order	9	17
Typography	7	4
Entity	6	0
Terminology	3	14
Addition	3	3
Grammar	2	1
Unintelligible	2	5
Untranslated	2	3
Character Encoding	1	0
Date/time	1	0
False Friend	1	0
Inconsistency	1	0
Number	1	0
Should not be tr-ed	1	0
Style/Register	1	6
Unidiomatic	1	0
Fluency	0	7

Table 1: Distribution of error types in Spanish and German corpora.

	PE time	PEE
Spanish	0.38	0.53
German	0.16	0.36

Table 2: Annotator agreement (Intraclass Correlation Coefficient)

Acknowledgements

Anna Zaretskaya is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement no317471.

References

Avramidis, E., Burchardt, A., Federmann, C., Popovic M., Tscherwinka C., & Vilar-Torres D. Involving Language Professionals in the Evaluation of Machine Translation. In Proceedings of the 8th ELRA Conference on Language Resources and Evaluation. Istanbul, Turkey, May 2012. 1127-1130.

Burchardt, A., Lommel, A. & Popovic, M. Deliverable 1.2.1. TQ Error Corpus. 31 July 2013.

Daems, J., Vandepitte, S., Hartsuiker, R. & Macken, L. The Impact of Machine Translation Error Types on Post-Editing Effort Indicators. In Proceedings of the 4th Workshop on Post-Editing Technology and Practice (WPTP4). Miami (Florida), October 30–November 3 2015. 31-45.

Federico, M., Cattelan, A. & Trombetti, M. Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA). 2012.

Koby, G. S. Editor's Introduction. In Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes. 2001. 1-23.

Koponen, M. Comparing human perceptions of post-editing effort with post-editing operations. In Proceedings of the 7th Workshop on Statistical Machine Translation. Montréal, Canada: Association for Computational Linguistics, June 2012. 181-190.

Koponen, M., Aziz, W., Ramos, L. & Specia, L. Post-editing time as a measure of cognitive effort. In Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012). San Diego, California, October 2012.

Krings, H. P. Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes. 2001.

Lacruz, I., Denkowski, M. & Lavie, A. Cognitive Demand and Cognitive Effort in Post-Editing. In Proceedings of the Third Workshop on Post-editing Technology and Practice. Vancouver (Canada), 26 October 2014.

Laübli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M. & Volk, M. Assessing Post-Editing Efficiency in a Realistic Translation Environment. In Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice. 2013. 83-91.

Lommel, A. Deliverable D 1.1.2. Multidimensional Quality Metrics. 28 June 2013.

Parra Escartín,C. & Arcedillo,M. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In Proceedings of the MT Summit XV, Research Track. Miami (Florida), November 2015. 131-144.

Popovic, M., Lommel, A., Burchardt, A., Avramidis, E. & Uszkoreit, H. Relations between different types of post-editing operations, cognitive effort and temporal effort. In proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 14). Dubrovnik, Croatia, 2014. 191-198.

Shrout, P. E. & Fleiss, J. L. Intra class Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin. 1979. 86, 2. 420-428.

Specia, L. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11). Leuven, Belgium, May 2011. 73-80.

Temnikova, I. Cognitive evaluation approach for a controlled language post-editing experiment. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta, May 2010.

Valotkaite, J. & Asadullah, M. Error Detection for Post-editing Rule-based Machine Translation. In Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012). San Diego, California, October 2012.

Zampieri, M. & Vela, M. Quantifying the Influence of MT Output in the Translators' Performance: A Case Study in Technical Translation. In Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat). May 2014. 93-98.

Zhechev, V. Analysing the post-editing of machine translation at Autodesk. In Post-editing of Machine Translation: Processes and Applications. Newcastle upon Tyne: Cambridge Scholars Publishing, 2014. 2-24.