Relations between different types of post-editing operations, cognitive effort and temporal effort

Maja Popović, Arle Lommel, Aljoscha Burchardt,
Eleftherios Avramidis, Hans Uszkoreit
DFKI – Berlin, Germany
name.surname@dfki.de

Abstract

Despite the growing interest in and use of machine translation post-edited outputs, there is little research work exploring different types of post-editing operations, i.e. types of translation errors corrected by post-editing. This work investigates five types of post-edit operations and their relation with cognitive post-editing effort (quality level) and post-editing time. Our results show that for French-to-English and English-to-Spanish translation outputs, lexical and word order edit operations require most cognitive effort, lexical edits require most time, whereas removing additions has a low impact both on quality and on time. It is also shown that the sentence length is an important factor for the post-editing time.

1 Introduction and related work

In machine translation research, ever-increasing amounts of post-edited translation outputs are being collected. These have been used primarily for automatic estimation of translation quality. However, they enable a large number of applications, such as analysis of different aspects of post-editing effort. (Krings, 2001) defines three aspects: temporal, referring to time spent on post-editing, cognitive, referring to identifying the errors and the necessary steps for correction, and technical, referring to edit operations performed in order to produce the post-edited version. These aspects of effort are not necessary equal in various situations.

Since the temporal aspect is important for the practice, post-editing time is widely used for measuring post-editing effort (Krings, 2001; Tatsumi, 2009; Tatsumi et Roturier, 2010; Specia, 2011). Human quality scores based on the needed amount of post-editing are involved as assessment of the cognitive effort in (Specia et al., 2010; Specia, 2011). Using edit distance between the original and the post-edited translation for assessment of the technical effort is reported in (Tatsumi, 2009; Tatsumi et Roturier, 2010; Temnikova, 2010; Specia, 2011; Blain et al., 2011).

More details about the technical effort can be obtained by analysing particular edit operations. (Blain et al., 2011) defined these operations on a linguistic level as post-editing actions and performed comparison between statistical and rule-based systems. (Temnikova, 2010) proposed the analysis of edit operations for controlled language in order to explore cognitive effort for different error types – post-editors assigned one of ten error types to each edit operation which were then ranked by difficulty. In (Koponen, 2012) post-edit operations are analysed in sentences with discrepancy between the assigned quality score and the number of performed post-edits. In one of the experiments described in (Wisniewski et al., 2013) an automatic analysis of post-edits based on Levenshtein distance is carried out considering only the basic level of substitutions, deletions, insertions and TER shifts. These edit operations are analysed on the lexical level in order to determine the most frequent affected words. General user preferences regarding different types of machine translation errors are explored in (Kirchhoff et al., 2012) for English-Spanish translation of texts from publich health domain, however without any relation to post-editing task. (Popović and Ney(, 2011)

number of sentences	quality level				
	ok	edit+	edit	edit-	bad
fr-en 2011	323	1559	0	544	99
en-es 2011	31	399	0	550	20
en-es 2012	200	548	856	576	74

Table 1: Corpus statistics: number of sentences assigned to each of the quality levels.

describe a method for automatic classification of machine translation errors into five categories, but only using independent human reference translations, not post-edited translation outputs.

The aim of this work is to systematically explore the relations of five different types of edit operations with the cognitive and the temporal effort. To the best of our knowledge, such study has not yet been carried out. Classification of edit operations is based on the edit distance and is performed automatically, and human quality level scores are used as a measure of cognitive effort.

2 Method and data

Experiments are carried out on 2525 French-to-English and 1000 English-to-Spanish translated sentences described in (Specia, 2011) as well as 2254 English-to-Spanish sentences used for training in the 2013 Quality Estimation shared task (Callison-Burch et al., 2012). All translation outputs were generated by statistical machine systems. For each sentence in these corpora, a human annotator assigned one of four or five quality levels as a measure for the cognitive effort:

- acceptable (ok)
- almost acceptable, easy to post-edit (edit+)
- possible to edit (edit)
- still possible to edit, better than from scratch (edit-)
- very low quality, better to translate from scratch than try to post-edit (bad)

Numbers of sentences assigned to each quality level are presented in Table 1.

All sentences were post-edited by the same two human translators[1] which were instructed to perform the minimum number of edits necessary to make the translation acceptable. Post-editing time is measured on the sentence level in a controlled way in order to isolate factors such as pauses between sentences.

The technical effort is represented by following five types of edit operations:

- correcting word form
- correcting word order
- adding omission
- deleting addition
- correcting lexical choice

The performed edit operations are classified on the word level using the Hjerson automatic tool (Popović, 2011) for error analysis. The post-edited translation output was used as a reference translation, and the results are available in the form of raw counts and edit rates for each category. Edit rate is defined as the raw count of edited words normalised over the total number of words i.e. sentence length of the given translation output.

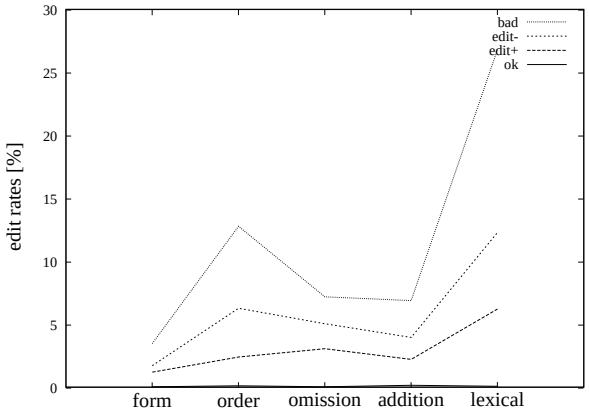
3 Results

3.1 Edit operations and quality level

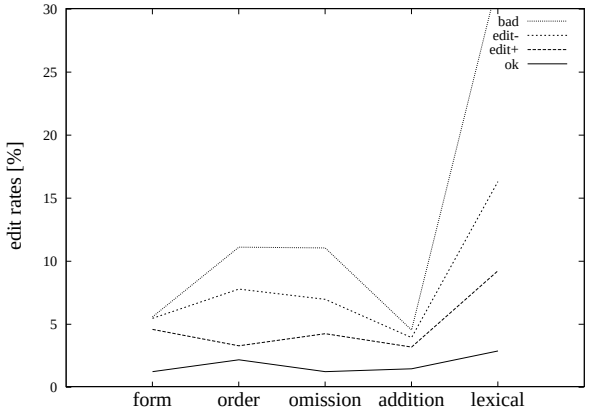
The distributions of five edit rates for different quality levels are presented in Figure 1. All edit rates increase with the decrease of quality, lexical choice and word order being the most prominent. The main difference between two edit types is that the number of lexical edits increases monotonically whereas the number of reordering edits is relatively low for high quality translations and relatively high for low quality translations.

Impact of reordering distance: In addition to five basic error types, we analysed reordering distances, i.e., the number of word positions by which a particular word is shifted. Reordering distances for different quality levels are presented in Figure 2. It can be seen that the distant reorderings are not an important issue, even for low quality translations, whereas the number of local and longer range reorderings both increase as quality decreases. The increase of longer ones, however, is more prominent for the low-quality translations: this relationship means that the increase of overall reordering errors presented in Figure 1 is primarily due to these reorderings. It should be noted that the experiments were carried out only on the language

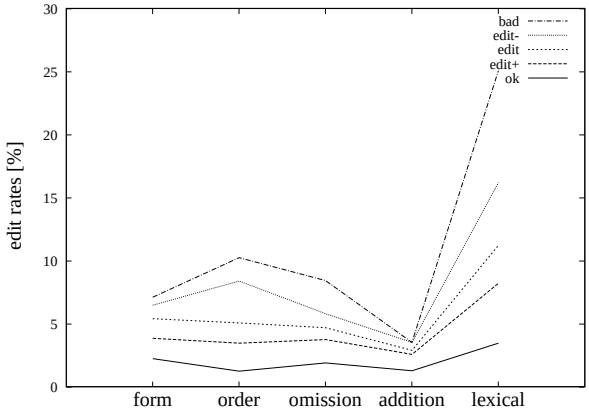
[1] One for French-English and one for English-Spanish output.



(a) French→English 2011

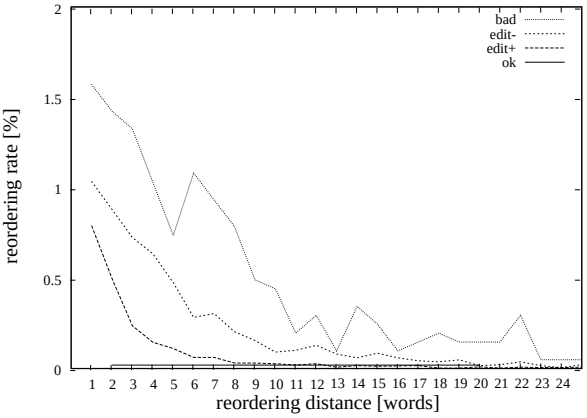


(b) English→Spanish 2011

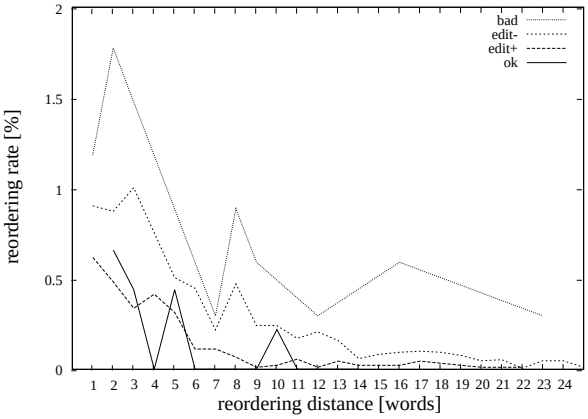


(c) English→Spanish 2012

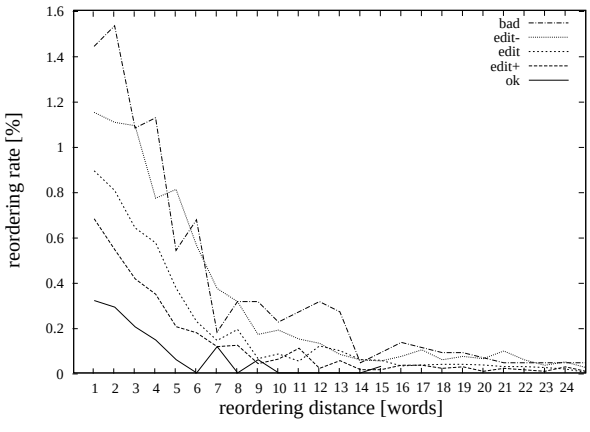
Figure 1: Distribution of five edit types for different quality levels in (a) one French-to-English and (b) two English-to-Spanish translation outputs.



(a) French→English 2011



(b) English→Spanish 2011



(c) English→Spanish 2012

Figure 2: Distribution of reordering distances for different quality levels in (a) one French-to-English and (b),(c) two English-to-Spanish translation outputs.

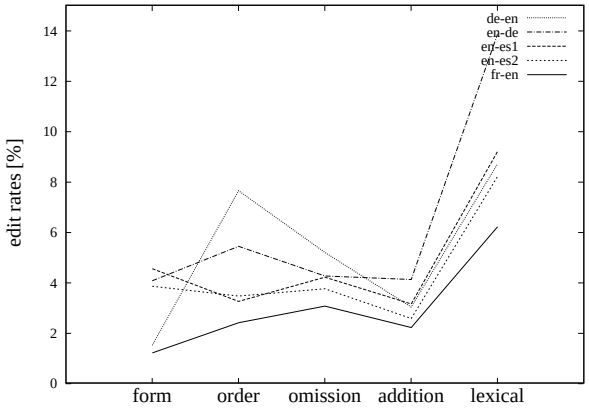
pairs with prevailing local structure differences – future experiments should include languages with different structure, such as German.

3.1.1 Almost acceptable translations

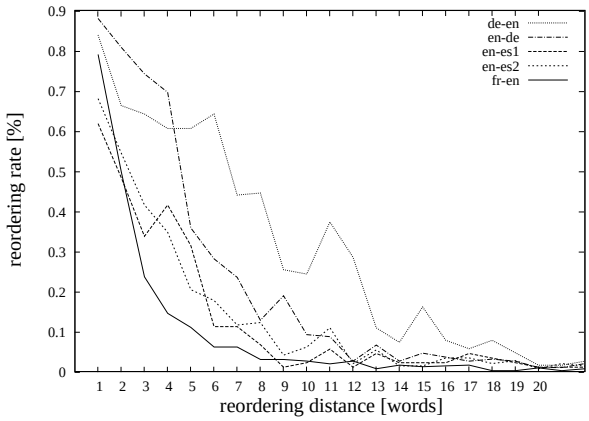
In addition to exploring different quality levels, we carried out an analysis only on almost acceptable translations for different language pairs. Almost acceptable translations are of the special in-

terest for high-quality machine translation – they are namely close to perfect translations and do not require much post-editing effort. The main question is which types of errors are keeping these translations from perfect.

For analysis of almost acceptable translations, apart from the sentences assigned to the "edit+" category in Table 1, an additional corpus was available, namely a portion of the German-to-English (778 sentences) and English-to-German (955 sentences) translations obtained by the best ranked statistical and rule based systems in the framework of the 2011 shared task (Callison-Burch et al., 2011).



(a) edit operations in almost acceptable translations



(b) reordering distances in almost acceptable translations

Figure 3: Distribution of (a) five edit operations and (b) reordering distances in almost acceptable translations: French-to-English, two English-to-Spanish, German-to-English and English-to-German outputs.

Distributions of five edit types as well as reordering distances in five almost acceptable sets are shown in Figure 3 and it can be seen that they are largely dependent on language pair and translation direction. The lexical edits are the most prominent for all translation directions indicating that even in the high-quality translations, large portions of texts are mistranslated. Inflectional errors are rare in high-quality English outputs, but still relatively high in Spanish and German translations. As for reordering errors, for French-English and English-Spanish translations the reordering edit rates are low, less than 4%, however for German-to-English translations it is almost 8% being not much lower than the lexical edit rate. This high rate indicates that, for this translation direction, even high-quality translations contain a significant number of syntactic errors. English-to-German, conversely, is quite difficult in general and the reordering edit rate is comparable to the rates for other types of operations; since all the edit rates are similar, improving any of them should lead to quality increase. As for reordering distances, short range reorderings are dominant in all high-quality translations, and the main difference for German-to-English outputs is due longer range reordering edits. Further analysis (e.g. based on POS tags) is needed to determine exact nature of reordering problems in the high quality translations.

3.2 Edit operations and post-editing time

Post-editing times are available for the 2011 data (first two rows in Table 1). The post-editing times for the English output are much shorter than for the Spanish output, probably due to language differences and/or to the different annotators. In any case, this difference does not represent an issue for estimating distribution of post-editing time over five edit operation classes. For each edit operation type, average post-editing time is calculated in the following way:

- for each sentence, divide the raw count of each edit type by the total number of edit operations thus obtaining weights;
- for each edit type in the sentence, estimate its post-editing time by multiplicating its weight with the whole sentence post-editing time;
- finally, for each edit type average the post-editing time over all sentences.

It should be noted that using uniform weights might be debatable on the sentence level but is sufficiently reliable on the document level. For example, if one sentence contains two lexical errors and

one word order error and the editing took 30 seconds, the estimated time for correcting each error type in this sentence is 10 seconds. However, it is theoretically possible that the reordering error actually took 20s and each of the lexical errors took only 5s. Nevertheless, many other sentences with different error distributions will be able to reflect this correctly. Therefore, averaging over all sentences gives a good estimate of post-editing time distribution over edit types. Distribution of post-editing time over reordering distances is calculated in a similar way, and all the results are presented in Figure 4.

It can be seen that the lexical edits require the largest portion of the time for both outputs. For the English translation output, the shortest time is needed for correction of the word form, and the times for other three edit types are similar. For the Spanish output, the deletion of extra words requires much less time than other edit types. As for reordering distances, as expected, longer reorderings require more time.

3.3 Quality level and post-editing time

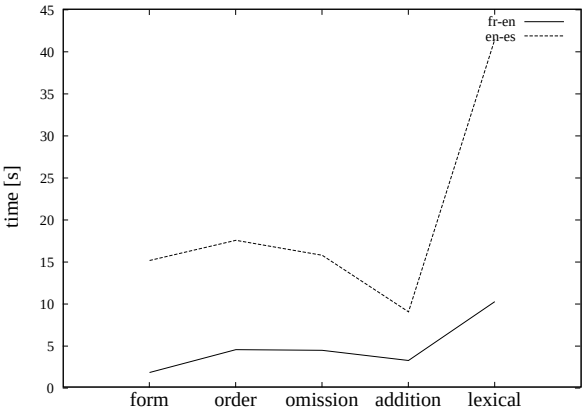
In previous sections, we compared five edit operation types with cognitive effort and with temporal effort separately. Nevertheless, the relation between these two aspects in the given context is also important to better understand all effects.

Post-editing times for different quality levels for the 2011 data are presented in Figure 5. Although an overall increase of the post-editing time can be observed when quality level decreases (i.e. cognitive effort increases), there is a discrepance for a significant number of sentences, especially for the sentences with low quality level score. In order to explore the reasons for differences between cognitive and temporal effort, further analysis of edit operations is carried out taking into accout both quality level and post-editing time.

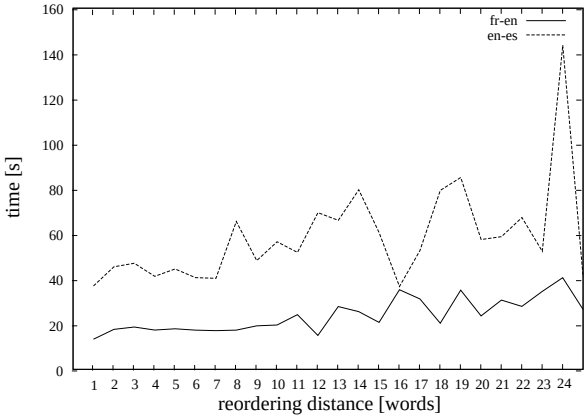
3.4 Analysis of discrepances

In order to examine differences between the cognitive and the temporal effort, we divided the texts in four parts:

- create two quality subsets: high-quality (edit+ and ok) and low-quality (edit- and bad) sentences
- calculate median post-editing time for low-quality sentences (which is 40 seconds for the English and 100 seconds for the Spanish output) and use it as a threshold
- create two time subsets for both quality subsets according to this threshold: “short-time” and “long-time” sentences.



(a) average post-editing time for five edit operations



(b) average post-editing time for different reordering distances

Figure 4: Average post-editing time (a) for five types of edit operations and (b) for different reordering distances: French-to-English and English-to-Spanish translation outputs.

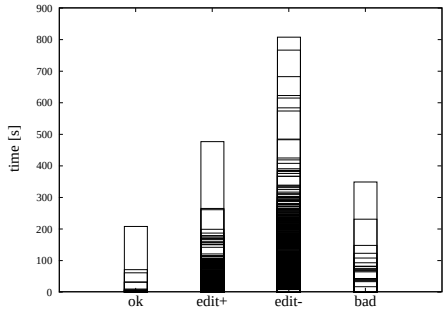
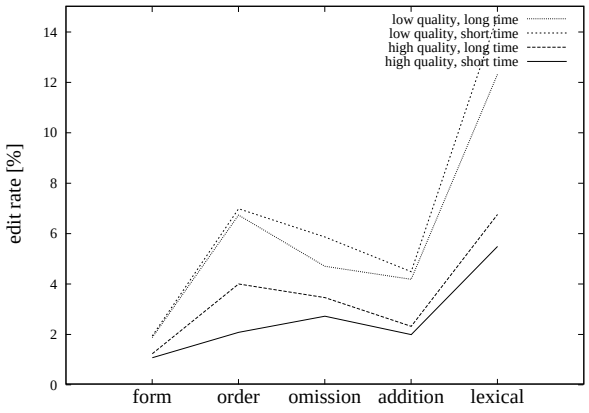
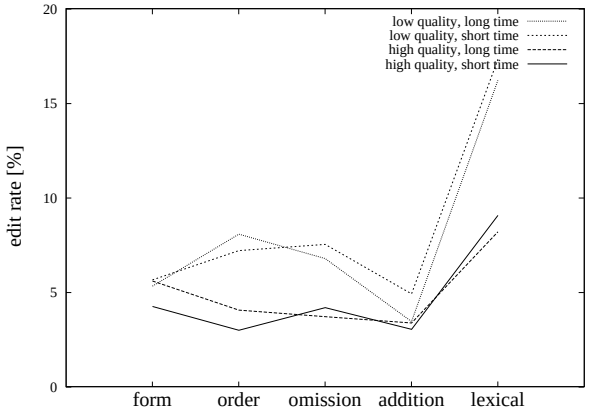


Figure 5: Distribution of post-editing times for different quality levels.

As a first step, edit rates for each subset are calcuated and the results are shown in Figure 6. The distributions for the same quality are very close – all edit rates are higher for the low-quality sentences regardless of the post-editing time. This indicates that the cognitive effort is tightly related to the amount of particular translation errors, mainly lexical and reordering errors, as already stated in Section 3.1.



(a) French→English 2011

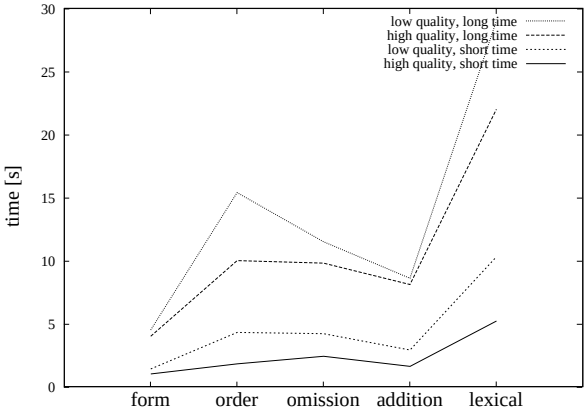


(b) English→Spanish 2011

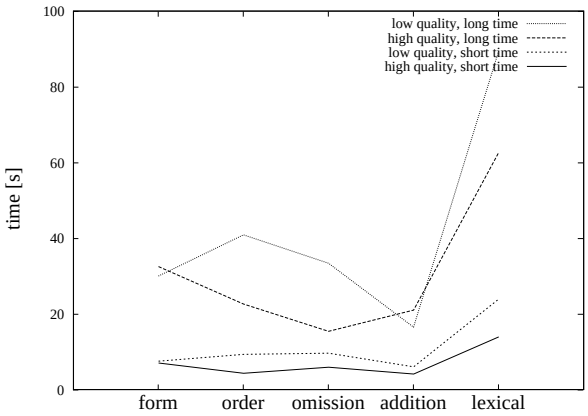
Figure 6: Edit rates for five edit operations – analysing discrepances between quality and time; (a) French-to-English and (b) English-to-Spanish output.

The next step was the analysis of post-editing time – what are the causes of long post-editing time for high quality translations and short post-editing time for low quality translations? For each sentence subset, average time distributions over five edit operation types are calculated as described in Section 3.2 and presented in Figure 7. The same tendencies can be observed for both translation outputs:

- all edit types required significantly more time in the long-time sentences than in the short-time sentences regardless of the quality level;
- low-quality translations required more time than high-quality translations in the same time subset;
  - this effect is larger for the long-time sentences,
  - especially for reordering errors, omissions and lexical corrections.



(a) French→English 2011



(b) English→Spanish 2011

Figure 7: Average post-editing times for five edit operations – analysing discrepances between quality and time; (a) French-to-English and (b) English-to-Spanish output.

The results confirm that the lexical and reordering errors require more post-editing effort than the others. In addition, post-editing time for low-quality translations is also affected by omissions, whereas this class has no significant importance in the high-quality translations.

These results also indicate the importance of the sentence length for the post-editing time (which

has also been observed in other studies, e.g. (Tatsumi, 2009; Koponen, 2012)). Edit rates are namely raw counts of edit operations normalised over the sentence length: since there is no significant variation of edit rates between the long-time and the short-time subset, the only remaining factor is the sentence length. On the other hand, a number of high-quality sentences require long post-editing time despite of low edit rates: the possible reason is that those sentences are longer.

In order to confirm this assumption, average sentence lengths were calculated for each sentence subset and the results are given in Table 2. As expected, long-time sentences are longer than short-time sentences regardless of the quality level. In addition, the relations of the sentence length with post-editing time and with quality level are presented in Figure 8: the post-editing time increases almost linearly with the increase of the sentence length, whereas the correspondence between the sentence length and the quality level is not straightforward, mainly due to the large number of short low-quality sentences.

quality	time	fr-en	en-es
high	short	22.7	19.6
	long	43.2	31.4
low	short	21.2	19.0
	long	40.6	35.5

Table 2: Average sentence lengths for four sentence subsets based on different quality levels and post-editing times.

4 Summary and outlook

We presented an experiment aiming to explore the relations of five different types of post-edit operations with the cognitive and the temporal post-editing effort. We performed automatic analysis of edit operations for different quality levels and estimated post-editing time for each of the five categories. The results showed that the reordering edits (shifts) and correcting mistranslations correlated most strongly with quality level i.e. cognitive effort, as well as that the lexical errors require the largest portion of post-editing time. Analysis of reordering distances showed that longer range reorderings have more effects both to the quality level and to the post-editing time, however very long ranges do not represent an issue.

In addition, we analysed the edit operations and reordering distances in almost acceptable translations in order to investigate which error types are present in almost perfect high-quality translations preventing them to be completely perfect. It is shown that the error distributions are dependent on the language pair and the translation direction: however, mistranslations are the dominant error type for all translation outputs.

Furthermore, we showed that the edit rates, especially for mistranslations and reorderings, correlate strongly with quality level regardless of the time spent on post-editing. On the other hand, post-editing time strongly depends on the sentence length.

Our experiment offers many directions for future work. First of all, it should be kept in mind that the French-English and English-Spanish language pairs are very similar in the terms of structure and morphology – word order differences are mostly of the local character, and both French and Spanish morphologies are rich mostly due to verbs. In future work, languages with more distinct structural differences (such as German) and richer morphology (such as Czech or Finnish) should be analysed. Furthermore, more details about edit operation types can be obtained by the use of additional knowledge such as POS tags.

Acknowledgments

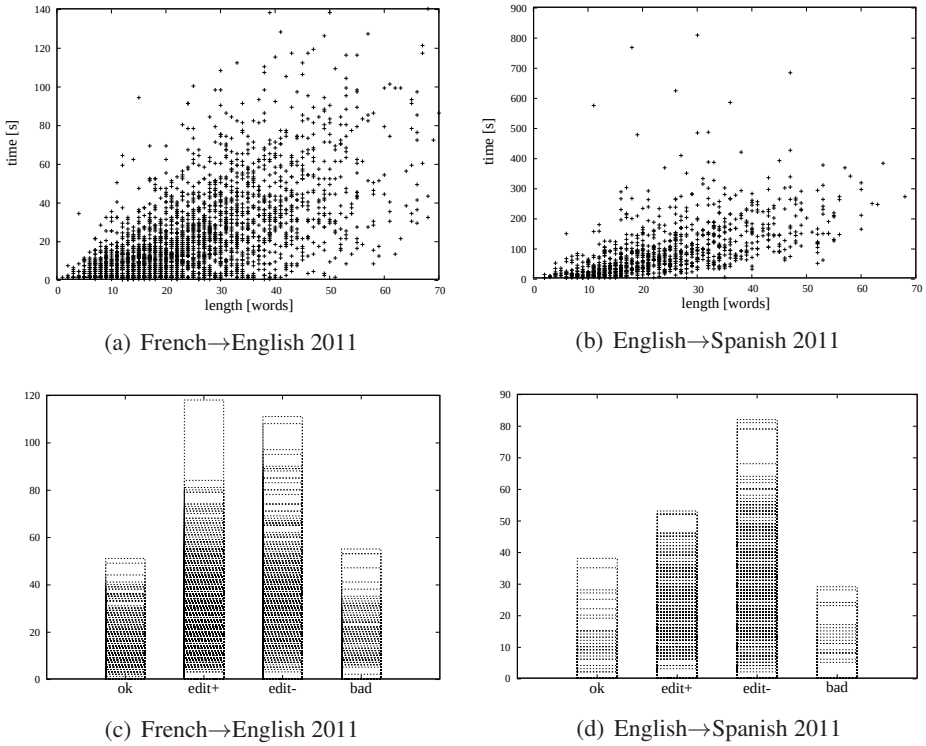
This work has been supported by the project QTLAUNCHPAD (EU FP7 CSA No. 296347).

References

Blain, Frédéric, Jean Senellart, Holger Schwenk, Mirko Plitt, and Johann Roturier. 2011. Qualitative analysis of post-editing for high quality machine translation. In *Machine Translation Summit XIII*, Xiamen, China, September.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 22–64, Edinburgh, Scotland, July.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 1051, Montral, Canada, June. Association for Computational Linguistics.



(a) French→English 2011

(b) English→Spanish 2011

(c) French→English 2011

(d) English→Spanish 2011

Figure 8: Distribution of post-editing times for (a),(b) different sentence lengths and (c),(d) different quality levels; (a),(c) French-to-English and (b),(d) English-to-Spanish output.

Kirchhoff, Katrin, Daniel Capurro, and Anne Turner. 2012. Evaluating user preferences in machine translation using conjoint analysis. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 12)*, pages 119–126, Trento, Italy, May.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190, Montral, Canada, June. Association for Computational Linguistics.

Krings, Hans. 2001. Repairing texts: empirical investigations of machine translation post-editing processes. Kent, OH. Kent State University Press.

Popović, Maja. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, (96):59–68, October.

Popović, Maja and Hermann Ney. 2011. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37(4), pages 657–688, December.

Specia, Lucia, Nicola Cancedda, and Marc Dymetman. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'2010)*, pages 3375–3378, Valletta, Malta, May.

Specia, Lucia. 2011. 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)*, pages 73–80, Leuven, Belgium, May.

Tatsumi, Midori. 2009. Correlation between automatic evaluation metric scores, post-editing speed and some other factors. In *Proceedings of MT Summit XII*, pages 332–339, Ottawa, Canada, August.

Tatsumi, Midori, Roturier, Johann. 2010. Source text characteristics and technical and temporal post-editing effort: what is their relationship?. In *Proceedings of the Second Joint EM+/CGNL Worskhop Bringing MT to the user (JEC 10)*, pages 43–51, Denver, Colorado, November.

Temnikova, Irina. 2010. 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.

Wisniewski, Guillaume, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and analysis of a large corpus of post-edited translations: quality estimation, failure analysis and the variability of post-edition. In *Proceedings of the MT Summit XIV*, pages 117–124, Nice, France, September.