# Comparing human perceptions of post-editing effort with post-editing operations

**Maarit Koponen**
University of Helsinki, Dept of Modern Languages
PO Box 24
00014 University of Helsinki, Finland
maarit.koponen@helsinki.fi

## Abstract

Post-editing performed by translators is an increasingly common use of machine translated texts. While high quality MT may increase productivity, post-editing poor translations can be a frustrating task which requires more effort than translating from scratch. For this reason, estimating whether machine translations are of sufficient quality to be used for post-editing and finding means to reduce post-editing effort are an important field of study. Post-editing effort consists of different aspects, of which temporal effort, or the time spent on post-editing, is the most visible and involves not only the technical effort needed to perform the editing, but also the cognitive effort required to detect and plan necessary corrections. Cognitive effort is difficult to examine directly, but ways to reduce the cognitive effort in particular may prove valuable in reducing the frustration associated with post-editing work. In this paper, we describe an experiment aimed at studying the relationship between technical post-editing effort and cognitive post-editing effort by comparing cases where the edit distance and a manual score reflecting perceived effort differ. We present results of an error analysis performed on such sentences and discuss the clues they may provide about edits requiring great cognitive effort compared to the technical effort, on one hand, or little cognitive effort, on the other.

## 1 Introduction

An increasingly common use for machine translation is producing texts to be post-edited by translators. While sufficiently high-quality MT has been shown to produce benefits for productivity, a well-known problem is that post-editing poor machine translation can require more effort than translating from scratch. Measuring and estimating post-editing effort is therefore a growing concern addressed by Confidence Estimation (CE) (Specia, 2011).

Time spent on post-editing can be seen as the most visible and economically most important aspect of post-editing effort (Krings, 2001); however, post-editing effort can be defined and approached in different ways. Krings (2001) divides post-editing effort into three types: 1. temporal, 2. cognitive and 3. technical. Temporal effort refers to post-editing time. Cognitive effort involves identifying the errors in the MT and the necessary steps to correct the output. Technical effort then consists of the keystrokes and cut-and-paste operations needed to produce the post-edited version after the errors have been detected and corrections planned. These different aspects of effort are not necessarily equal in various situations. In some cases, the errors may be easy to detect but involve several technical operations to be corrected. In other cases, parsing the sentence and detecting the errors may require considerable cognitive effort, although the actual technical operations required are quick and easy. According to Krings (2001), temporal effort is a combination of both cognitive and technical effort, with cognitive effort being the decisive factor. Assessing and reducing the cognitive effort involved in MT post-editing would therefore be important but the task is far from simple. Past experiments have involved cognitive approaches such as think-aloud protocols (Krings, 2001; O'Brien, 2005; Carl et al., 2011) and

*Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 181–190,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

post-editing effort scores assigned by human evaluators (Specia et al., 2009; Specia, 2011; Specia et al., 2011).

While edit operations reflect the amount of technical effort needed, subjective assessments of perceived post-editing effort needed can serve as a measure of cognitive post-editing effort: in order to give such an estimate, the evaluator needs to cognitively process the segment in order to detect the errors and plan the necessary corrections. Using these two measures, a comparison of technical effort and perceived amount of post-editing effort can serve as a way to evaluate cognitive post-editing effort. We propose that studying cases where the perceived effort necessary is greater or smaller than the number of actual edit operations performed may provide clues to situations where the cognitive and technical effort differ. Cases where the human editor overestimates the need for editing (as compared to number of edit operations performed) could indicate that these segments contain errors requiring considerable cognitive effort. On the other hand, cases where the manual score underestimates the amount of editing needed could indicate errors that require relatively little cognitive effort compared to the number of technical operations.

To examine the question of differences in technical and cognitive post-editing effort, we present an analysis of MT segments that have different levels of post-editing indicated by the manual effort score and actual number of post-edit operations indicated by the edit distance. By analyzing cases where these two measures of post-editing effort differ, it may be possible to isolate cases that require more cognitive effort than technical effort and vice versa. Section 3 describes the material and method used in the experiment, and the results of the analysis are presented in Section 4.

## 2 Related work

As the temporal aspect of post-editing effort is important for the practice of machine translation post-editing, post-editing time has been a commonly used measure of post-editing effort (Krings, 2001; O'Brien, 2005; Specia et al., 2009; Tatsumi, 2009; Tatsumi and Roturier, 2010; Specia, 2011; Carl et al., 2011). The technical aspect of post-editing effort has been approached by following keystrokes and cut-and-paste operations (Krings, 2001; O'Brien, 2005; Carl et al., 2011) or using automatic metrics for edit distance between the raw MT and post-edited version (Tatsumi, 2009; Temnikova, 2010; Tatsumi and Roturier, 2010; Specia and Farzindar, 2010; Specia, 2011; Blain et al., 2011). Several edit operations may also be incorporated in one "post-edit action (PEA)", introduced by Blain et al. (2011). For example, changing the number of a noun propagates changes to other words, such as the determiners and adjectives modifying it. Tatsumi and Roturier (2010) also explore the relationship between temporal and technical aspects of post-editing effort.

Cognitive aspects of post-editing effort have been approached with the help of keystroke logging (Krings, 2001; O'Brien, 2005; Carl et al., 2011) and gaze data (Carl et al., 2011), attempting to measure cognitive effort in terms of pauses and fixations. O'Brien (2005) also experiments with the use of choice network analysis (CNA) and think-aloud protocols (TAP). Human scores for post-editing effort have involved assessing the amount of post-editing needed (Specia et al., 2009; Specia, 2011) or adequacy of the MT (Specia et al., 2011).

Temnikova (2010) proposes the analysis of the types of changes and comparison to post-editing time as a way to explore cognitive effort. For this purpose, Temnikova (2010) builds upon the MT error classification by Vilar et al. (2006) and their own post-editing experiments using controlled language to draft a classification for the cognitive effort required for correcting different types of MT errors. This classification defines ten types of errors and ranks them from 1 to 10 with 1 indicating the easiest and 10 the hardest error type to correct. The easiest errors are considered to be connected to the morphological level, or correct words with incorrect form, followed by the lexical level, involving incorrect style synonyms, incorrect words, extra words, missing words and erroneously translated idiomatic expressions. The hardest errors in the classification relate to syntactic level and include wrong punctuation, missing punctuation, then word order at word level and finally word order at phrase level. The ranking is based on studies in written language comprehension and error detection. Results reported in Temnikova (2010) suggest that pre-edited machine translations that had previously been found to require less post-editing effort measured by post-edit time and edit distance contain less errors that are cognitively more difficult compared to MT that had not been pre-edited.

In this study, we aim to investigate the relationship between the cognitive effort and the technical effort involved in post-editing. Edit distance between MT segments and their post-edited versions is used as a measure of technical effort and human effort scores as a measure of cognitive effort.

## 3 Material and method

The data used in this study consists of English to Spanish MT segments from the evaluation task training set provided for the quality estimation task at the NAACL 2012 Seventh Workshop on Statistical Machine Translation WMT12. [1] The training set consists English to Spanish machine translations of news texts, produced by a phrase-based SMT system. The data available for each segment includes the English source segment, Spanish reference translation produced by a human translator, machine translation into Spanish, post-edited version of the machine translation and a manual score indicating how much editing would be required to transform the MT segment into a useful translation. The manual score included is the average of scoring conducted by three professional translators using a 5-point scale where (1) indicates the segment is incomprehensible and needs to be translated from scratch, (2) significant editing is required (50-70% of the output), (3) about 25-50% of the output needs to be edited, (4) about 10-25% needs to be edited, and (5) little to no editing is required.

Additional information includes the SMT alignment tables. The alignments were not part of the original set, and in some cases differed slightly from the segments that had been used for the manual scoring. As we intended to make use of the alignments from source to MT, we included only segments that were identical in the original evaluated set.

To measure the amount of editing performed on the segments, the translation edit rate (TER) (Snover et al., 2006) was calculated using the post-edited versions as reference. TER measures the minimum number of edits that are needed to transform the machine translation into the post-edited segment used as reference. Edits can be insertion, deletion, substitution or reordering and the score is calculated as the number of edits divided by the number of tokens in the reference. The higher the TER score, the more edits have been performed.

As our aim was to focus on cases where the perceived effort score and the amount of editing differed, we looked for two types of sentences at the opposite ends of the manual effort scoring scale: (1) Cases where the manual score indicated more editing was needed than had actually been performed. (2) Cases where the manual score indicated less editing was needed than had actually been performed.

For Case (1), we selected segments with a manual score of 2.5 or lower, meaning that at least 50% of the segment needed editing according to the evaluators. We looked for the ones with the lowest TER scores, trying to find at least 30 sentences. The set selected for analysis consists of 37 sentences with a manual effort score of 2.5 or lower and TER score 0.33 or lower. For comparison, we also selected the same number of sentences with similar TER scores but with manual scores of 4 or above. These sets are referred to as the low TER set.

For Case (2), we selected segments with a manual score of 4 or above, meaning that no more than 25% of the segment needed editing according to the evaluator. Again, we looked for about 30 sentences with the highest TER scores. The set selected consists of 35 sentences with a manual effort score of 4 or higher, and TER score 0.45 or higher. For comparison, we also selected sentences with similar TER scores but low manual scores. These sets are referred to as the high TER set.

The selected MT segments and post-edited versions were then tagged with the FreeLing Spanish tagger (Padró et al., 2010). The tagged versions contain the surface form of the word, lemma and a tag with part-of-speech (POS) and grammatical information. Other tools such as dependency parsing were considered, but within the scope of this study, we decided to experiment what changes can be observed using only the basic lemma, POS and form information.

The tagged versions were aligned manually, first

[1] http://www.statmt.org/wmt12/quality-estimation-task.html

matching identical tokens (words and punctuation) in the sentence, then matching words with the same lemma but different surface form. The alignment table was consulted to match substitutions that involved a different word and even different POS. Each matched pair of words in the MT and post-edited versions was then labeled to indicate whether the match was identical or involved editing the word form, substituting with a different word of the same POS or a word of different POS. Words appearing in the post-edited version but not in the MT were labeled as insertions and words appearing in the MT but not in the post-edited version as deletions. In cases where several MT words were replaced with one in the post-edited version or one MT word was replaced with many in the post-edited, a match was made between words of the same POS and form, if such was found, or the first word in the sequence if none matched. The remaining words were labeled as inserted/deleted.

The positions of the matched words were also compared. For matching the word order, changes caused only by insertion or deletion of other words were ignored, and words that had remained in the same order after post-editing were labeled as same. In cases where the word order did not match, the word was labeled with the distance it had been moved and whether it had been moved alone or as a part of a larger group.

The totals of changes within a sentence were then calculated and the patterns of changes made by editors were examined. In addition to the total number of edit operations, we considered the possibility that editing certain parts-of-speech might require more effort than others. In particular, editing content words such as verbs or nouns might require more effort than editing function words such as determiners, because they are more central to conveying the content of the sentence. Further, as Blain et al. (2011) argue, changes to these words may propagate changes to other words in the sentence. Punctuation was also treated separately to follow Temnikova's (2010) classification of punctuation errors as a class of their own.

The patterns found in the sample sentences were compared to the comparison sets of sentences with similar TER scores. Additionally, Spearman rank correlations between the manual effort score and the various edit categories were calculated for all tokens and specific POS classes. The next section presents the results of these comparisons.

## 4 Results

This section presents the results from the analysis of post-editing changes. The total number of segments and tokens and the percentages of edited and reordered tokens in each set are shown in Table 1. Comparisons of the edit patterns between segments with similar TER scores but different manual scores are shown in Figures 1 to 4. Figure 1 presents the distributions of edit categories in the low TER sets and Figure 3 in the high TER sets. Figure 2 presents the percentages of changed tokens and reordered tokens by POS class in the low TER set and Figure 4 in the high TER sets. In Figures 2 and 4, nouns, verbs, adjectives and determiners are shown separately, while other parts-of-speech are combined into "Other". Punctuation is also presented separately.

Tables 2 and 3 present Spearman rank correlations between the manual score and different edit categories. Overall correlations regardless of POS are given for all edit categories. For specific POS classes, only the edit categories with strongest correlations are listed in each case.

### 4.1 Case 1: Low TER set

These sentences represent a case where the human evaluators indicated that significant post-editing would be needed but the low TER score indicated that relatively little editing had been performed. The most noticeable difference between segments with high and low manual scores is the number of tokens: low-scored segments have about twice as many tokens on average than the high-scored ones (see Table 1) and the number of tokens in the post-edited segment has a strong negative correlation (Table 2). Besides segment length, other strong correlations involve different types of reordering. Reorderings involving a distance of one step show weaker correlation than changes involving a longer distance. No correlation was found for any of the word change categories in this case.

Broken down by the POS class, results are similar to the overall result in that reordering categories have the strongest (negative) correlations with the

| TER score | Manual score | Number of segments | Number of tokens | Edited tokens | Reordered tokens |
|---|---|---:|---:|---:|---:|
| Low | Low | 37 | 1480 | 23% | 24% |
| Low | High | 37 | 695 | 21% | 15% |
| High | Low | 35 | 943 | 45% | 45% |
| High | High | 35 | 556 | 42% | 33% |

Table 1: Total number of sentences and tokens per set, percentage of tokens edited and percentage of tokens reordered.

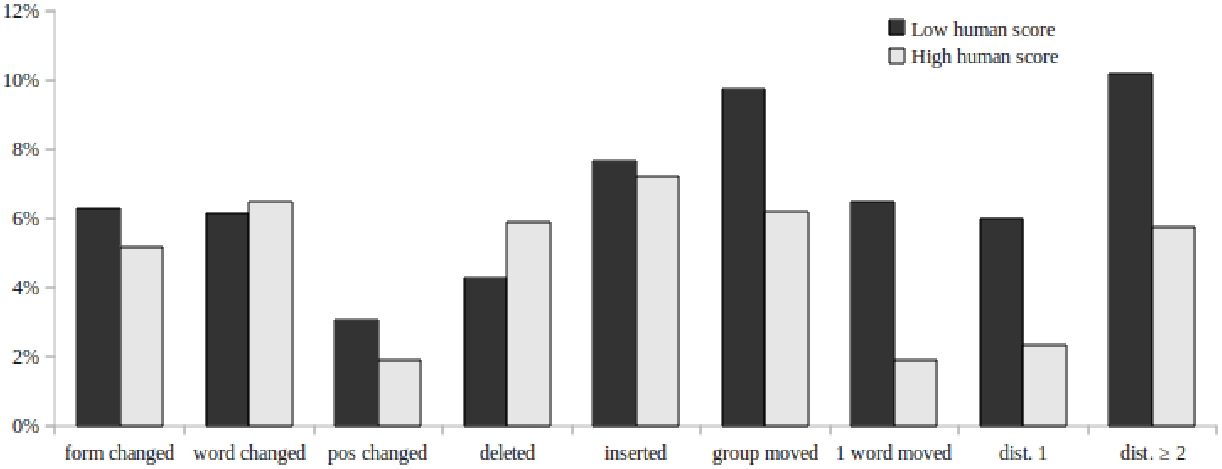
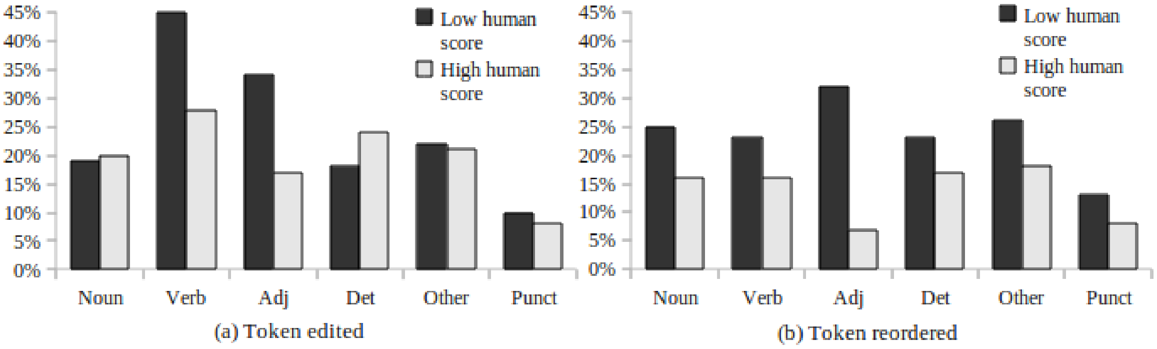


Figure 1: Distribution of edit categories - Low TER.



(a) Token edited

(b) Token reordered

Figure 2: Edited and reordered tokens by POS - Low TER

effort score. Strongest correlations also mostly involve nouns, adjectives or verbs. As shown in Figure 2, the differences in percentage of edited tokens are largest for verbs and adjectives. In high-scored sentences, 72% of verbs were unchanged by the editor compared to 55% in the low-scored ones. In both cases, most edits to verbs involved changing the form of the verb, (23% in low-scored vs 11% in high-scored). Adjectives have a similar pattern with 18% of edited adjective forms in low-scored vs 7% in high-scored sentences.

Sentences with high manual scores actually have more cases of edited determiners and nouns, although for nouns the difference is only 1%. Most edits to determiners involved deletion (15% of determiners) or changed form (11%) in the case of high-scored sentences. In low-scored sentences, insertion was most common (10% of determiners). Within

| **Overall correlations** | | |
|---|---|---|
| number of tokens | -0.51 | *** |
| word match | 0.11 | |
| form changed | -0.10 | |
| word changed | -0.15 | |
| pos changed | -0.15 | |
| deleted | 0.08 | |
| inserted | -0.15 | |
| order same | 0.51 | *** |
| group moved | -0.48 | *** |
| 1 word moved | -0.47 | *** |
| dist. 1 | -0.37 | ** |
| dist $\geq$ 2 | -0.53 | *** |
| **Strongest correlations by POS** | | |
| Noun, order same | 0.49 | *** |
| Adj, order same | 0.47 | *** |
| Noun, group moved | -0.46 | *** |
| Adj, dist. $\geq$ 2 | -0.46 | *** |
| Noun, dist. $\geq$ 2 | -0.45 | *** |
| Other, group moved | -0.44 | *** |
| Verb, 1 word moved | -0.44 | *** |
| Verb, dist. $\geq$ 2 | -0.43 | *** |
| Other, order same | 0.41 | *** |
| Det, group moved | -0.40 | *** |
| Verb, word match | 0.39 | *** |
| Adj, 1 word moved | -0.38 | *** |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 2: Spearman rank correlations between effort score and edit categories - Low TER.

the class ”Other” combining numbers, adverbs, conjunctions, pronouns and prepositions, adverbs were an similar case in that there were more unchanged adverbs in the low-rated sentences (86%) than in the high-rated (72%). However, the total number of adverbs in either set was very small.

### 4.2 Case 2: High TER set

These sentences represent a case where the human evaluators indicated only a little editing was needed but the high TER score indicated much more editing had been performed. Again one noticeable difference between the sentences with low and high manual scores is the number of tokens (see Table 1), although the negative correlation shown in Table 3 was not as strong as for the low TER set.

For these sentences, word changes have stronger correlations with the manual effort score (Table 3). While the shares of fully matched words are fairly equal between the sentences, differences appear in some of the edit categories. Sentences with high manual scores have more cases where the word form has been edited (Figure 3), and changed form has the strongest (positive) correlation after number of tokens. High-scored segments also appear to have more deletions, but essentially no correlation was found between the manual score and deletions on the segment level. As shown in Figure 3, low-scored segments have more cases of substitution with different word. Reordering is again more common in low-scored segments, but correlations for reordering are weaker than in the low TER set. Cases where one word has been moved alone rather than as a part of a group has the strongest correlation among the reordering categories.

| **Overall correlations** | | |
|---|---|---|
| number of tokens | -0.43 | *** |
| word match | 0.14 | |
| form changed | 0.36 | ** |
| word changed | -0.25 | * |
| pos changed | -0.28 | * |
| deleted | 0.14 | |
| inserted | -0.22 | |
| order same | 0.21 | |
| group moved | -0.12 | |
| 1 word moved | -0.34 | ** |
| dist. 1 | -0.22 | |
| dist. $\geq$ 2 | -0.25 | * |
| **Strongest correlations by POS** | | |
| Other, inserted | -0.38 | ** |
| Noun, 1 word moved | -0.36 | ** |
| Noun, pos changed | -0.35 | ** |
| Noun, word changed | -0.30 | * |
| Adj, order same | 0.28 | * |
| Det, inserted | -0.27 | * |
| Adj, dist. $\geq$ 2 | -0.25 | * |
| Noun, word match | 0.24 | * |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 3: Spearman rank correlations between effort score and edit categories - High TER.

For specific POS classes, the strongest correlation

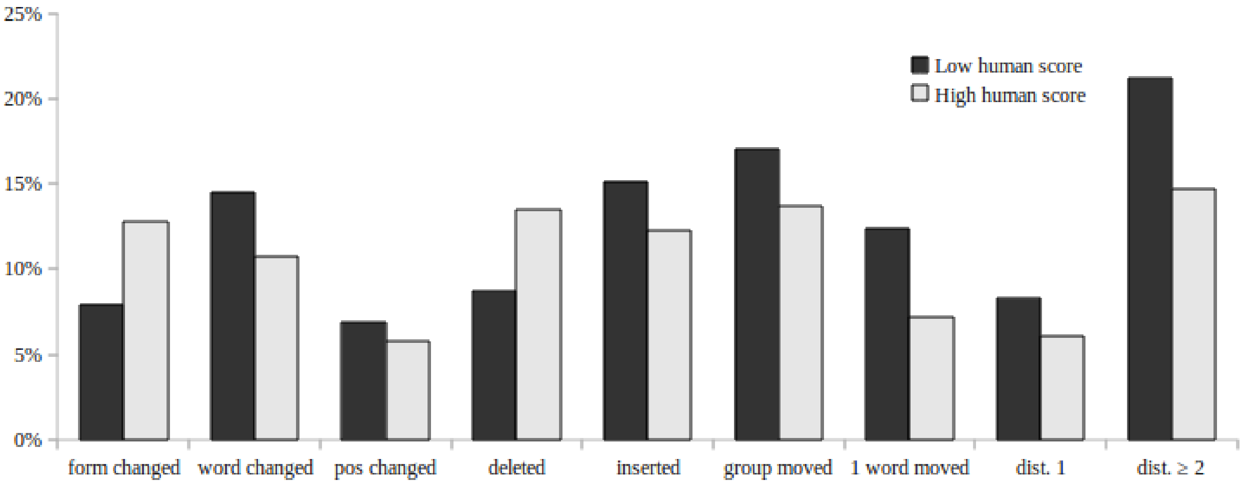
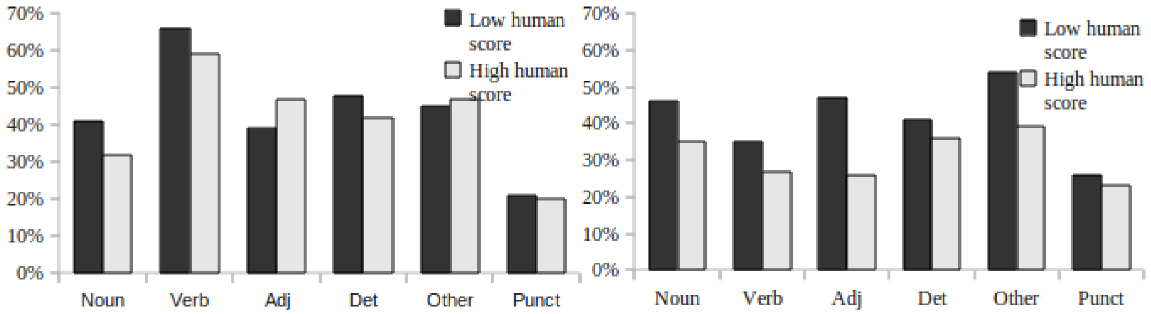


Figure 3: Distribution of edit types - High TER.



(a) Token edited

(b) Token reordered

Figure 4: Edited and reordered tokens by POS - High TER

in Table 3 involves insertion of words in the combined class ”Other” (numbers, adverbs, conjunctions, pronouns and prepositions). Within this class, pronouns actually required most edits: in low-scored segments, 50% of pronouns were inserted by the editor (32% in high-scored segments). The largest difference in the percentage of edited tokens is seen with nouns (41% edited in low-scored segments vs 32% in high-scored, and edits related to nouns are also among the strongest correlations for this set. In the case of adjectives, the segments with low manual score actually have more cases where no editing of the word has been required (61% vs 53%), but high-scored sentences contain a larger share of cases (32% vs 16%) where only the form of the adjective has been edited. However, these correlations remained weak. Reordering involving nouns and adjectives, on the other hand, again appears among the strongest correlations.

## 5 Discussion

Perhaps the most obvious difference between segments with high and low manual scores is segment length: long segments tend to get low scores even when the amount of editing turns out to be less than estimated. The effect of sentence length has also been observed in other studies, e.g. (Tatsumi, 2009). One simple explanation would be that a high total number of words leads to a high total number of changes to be made and therefore involves considerable technical post-editing effort. However, as the case of segments with low manual scores but low TER show, sometimes these long sentences do not, in fact, require a large number of edit operations.

This suggests also increased cognitive effort, as the sheer length may make it difficult for the evaluator/editor to perceive what needs to be changed and plan the edits.

We also noticed during the analysis that some of the very long segments actually consisted of two sentences. Furthermore, in some these cases, one of the sentences contained few changes while most of the changes were confined to the other. Similarly, long segments consisting of only one sentence sometimes contained long unchanged passages while some other part of the sentence was edited significantly. In these cases, such unchanged passages could be useful to the post-editor in real life situations, but the error-dense passage affects perception of the segment as a whole. Perhaps this suggests that assessing MT for post-editing and post-editing itself could benefit from presenting longer segments in shorter units, allowing the evaluator or editor to choose and discard different units within a longer segment.

Tatsumi (2009) also found that very short sentences increased post-editing time. In this study, all extremely short sentences found had received high scores from the human evaluators. Some are found in the low TER/high manual score set used for comparison purposes, but there are also some in the set of sentences with high TER/high manual score, meaning that there were relatively many edits compared to the length of the segment but the evaluators had indicated that little editing was needed. At least for the segments analyzed here, it appears that the evaluators did not consider short sentences to require much effort regardless of the actual number of edits performed. In Tatsumi's (2009) results, also other aspects, such as source sentence structure and dependency errors in the MT were discovered to have an effect on post-editing time. In this study, sentence structure and dependency errors were not explicitly examined, but these aspects would be of interest in future work.

Edits related to reordering also appear to be connected to low manual scores, as low-scored sentences involved more reordering than high-scored ones in both cases. This reflects Temnikova's (2010) error ranking where errors involving word order, particularly at phrase level, are considered the most difficult to correct. Besides the number of reorderings necessary, the results of this study may suggest some differences in whether reordering involves isolated words or groups of words and distances of one step (word level order) or longer distances.

Examining the results by parts-of-speech may suggest that overall, edits related to nouns, verbs or adjectives take more effort than other POS, because in both sets, strongest correlations mainly involved nouns, verbs and adjectives. In both sets, sentences with low manual scores contained more cases of edited verbs, and verb matches had one of the strongest correlations in the low TER set. On the other hand, edits related to nouns appeared to have particularly strong correlations in the high TER set. In this set, however, the strongest negative correlation was found for insertion of the other POS (mainly pronouns), so at least some of the other POS may also be difficult to edit.

Some cases where relatively little cognitive effort is required may be suggested by the situations where the high-scored sentences in fact contain more edits than the low-scored ones. In the high TER set, sentences with high manual scores contained more cases where only the form of a word has been edited, whereas sentences with low manual scores contained more cases of substitution with a different word or even different POS. This reflects the ranking of such errors in (Temnikova, 2010), where word form errors are considered cognitively easiest. This particularly appears to be the case for adjectives in this set. Although segments with a high manual score actually have a smaller number of fully correct adjectives than low-scored ones, they contain a larger share of instances where only the form of the adjective has been edited. Another example of edits involving less cognitive effort might be determiners in the low TER set, where again sentences with high manual scores contain more edited determiners than those with low scores. In this case, deletion of determiners was common in addition to changing the form.

Overall, deletion and insertion or extra words and missing words appeared to have little effect. While sentences with high manual scores have a slightly higher percentage of deleted words in both sets, the correlation was weak. Most of the deletions of content words seemed to involve auxiliary verbs, but in some instances it is difficult to say whether the editor has, in fact, considered something "extra" information and why, whether there has been a deliberate choice to implicitate certain information or whether the deletion has been at least partly unintentional. During the alignment process of the MT and post-edited version, it appeared that some source elements, in some cases entire clauses and in others certain words, were completely missing in the post-edited version. On the other hand, some of the insertions were also difficult to map onto anything in the source segment and the editor appeared to have brought in something extra. One clear example involved adding a conversion from miles per hour to km per hour that did not appear in the MT or source text. Such deletions and insertions concerned only a few isolated cases which were not examined in detail within the scope of this work. Some error classifications, such as Blain et al. (2011), do also take errors made by post-editors into account, and one interesting aspect of post-editing would be to study the correctness of post-edits. If it would turn out that post-editors are more prone to make errors or to fail to correct errors, (particularly errors related to content as opposed to typographical errors etc.) in certain situations, this might suggest situations that involve particular cognitive effort or mislead the editor.

## 6 Conclusion and Future Work

We have presented an experiment aimed at exploring the difference between cognitive and technical aspects of MT post-editing effort by comparing human scores of perceived effort necessary to actual edits made by post-editors. We examined cases where considerably more or considerably less post-editing was done than predicted by the evaluators' estimate of post-editing needed. The results show that one of the factors most affecting the perception of post-editing necessary involves segment length: long segments are perceived to involve much effort and therefore receive low scores even when the actual number of edits turns out to be small. This suggests that sentence length affects the cognitive effort required in identifying errors and planning the corrections, and presenting MT for this type of evaluation and post-editing may benefit from displaying segments to the evaluator or editor in smaller units.

The results also suggest other features affecting cognitive effort. Sentences with low manual scores were found to involve more reordering, indicating increased cognitive effort, while sentences with high manual scores were found to involve more cases of correct words with incorrect form, suggesting that these errors are cognitively easier. Examining edit type distributions in different POS classes suggests that edits related to certain parts-of-speech, namely nouns, verbs and adjectives, may also be associated with perception of more effort. On the other hand, sentences with high scores in some cases contained even more editing of some other POS and types, such as editing forms of adjectives or deleting determiners, which may indicate that these errors affect perception of effort to a lesser extent. As the number of sentences used was relatively low, however, such effects would require more study.

In future work, we aim to more explicitly examine combinations of edit operations, (e.g. changing the form and reordering, moving a group and substituting one word within the group) and features such as dependency errors (Tatsumi, 2009). Further experiments with data on other language pairs would also be needed. Another interesting aspect for future work would be trying to distinguish between edits made for reasons of incorrect language and edits for reasons of incorrect content. Further, examining the success of post-editing and exploring whether post-editors themselves are prone to make errors or fail to correct errors in certain situations could be an interesting avenue for discovering situations that involve significant cognitive effort.

## Acknowledgments

This work has been supported by LANGNET Finnish Graduate School in Language Studies.

## References

Frédéric Blain, Jean Senellart, Holger Schwenk, Mirko Plitt and Johann Roturier 2011. Qualitative analysis of post-editing for high quality machine translation. In *MT Summit XIII: the Thirteenth Machine Translation Summit* [organized by the] Asia-Pacific Association for Machine Translation (AAMT), pages 164-171. 19-23 September 2011, Xiamen, China.

Michael Carl, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt Lykke Jakobsen. 2011. The process of post-editing: a pilot study. In *Proceedings of* *the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation*, pages 131-142. Copenhagen Business School, 20-21 August 2011. (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur.

Hans P. Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing process*. The Kent State University Press, Kent, OH.

Sharon O'Brien 2005. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation*, 19(1):37-58.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes and Irene Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation*, pages 3485-3490. 17-23 May 2010, Valletta, Malta.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223-231. August 8-12, 2006, Cambridge, Massachusetts, USA.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman and Nello Cristianini 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the EAMT*, pages 28-35. Barcelona, May 2009.

Lucia Specia and Atefeh Farzindar. 2010. Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, pages 33-41. Denver, CO, 4 November 2010.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-Editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73-80. Leuven, Belgium, May 2011.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett and Wilker Aziz. 2011. Predicting Machine Translation Adequacy. In *MT Summit XIII: the Thirteenth Machine Translation Summit* [organized by the] Asia-Pacific Association for Machine Translation (AAMT), pages 513-520. 19-23 September 2011, Xiamen, China.

Midori Tatsumi. 2009. Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 332-339 August 26-30, 2009, Ottawa, Ontario, Canada.

Midori Tatsumi and Johann Roturier. 2010. Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship?. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, pages 43-51. Denver, CO, 4 November 2010.

Irina Temnikova. 2010. Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation*, pages 3485-3490. 17-23 May 2010, Valletta, Malta.

David Vilar, Jia Xu, Luis Fernando D'Haro and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings*, pages 697-702. Genoa, Italy, 22-28 May 2006.