Koponen, M., & Salmi, L. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 137–148.

# Post-editing quality: Analysing the correctness and necessity of post-editor corrections

**Maarit Koponen**

University of Turku, Finland
maarit.koponen@utu.fi

**Leena Salmi**

University of Turku, Finland
leena.salmi@utu.fi

*Post-editing (PE) machine translations (MT) has become an increasingly common practice in the translation field in recent years. Research has investigated, among other issues, the types of error corrected by post-editors, but less emphasis has been placed on the corrections themselves and how they reflect MT errors. This article presents a pilot study analysing the edits made by five student post-editors in an English–Finnish post-editing task. We analyse the correctness and necessity of the edits. Our results show that, whereas most edits performed in the task are correct, a significant number of them (34%) are unnecessary. The findings suggest that specific types of edit, such as word-order changes and deletions of personal pronouns, are generally unnecessary for this language pair, which may have implications for post-editing practice and training.*

## 1. Introduction

The use of machine translation (MT) and post-editing (PE) has increased considerably in the translation field in recent years and it has been shown that the practice can increase productivity without compromising the quality of the final translation (e.g., Plitt & Masselot, 2010). Research has also investigated the types of MT error corrected by post-editors and the relationship between errors and post-editing effort; it has also attempted to differentiate between errors affecting meaning and those affecting language (Daems, Vandepitte, Hartsuiker, & Macken, 2015; Koponen & Salmi, 2015; Lacruz et al., 2014; Popović, Lommel, Burchardt, Avramidis, Uszkoreit, 2014). Increasing evidence shows that, in addition to the number of errors edited, the type of errors affects PE effort, and some errors are more demanding than others (Koponen, 2012; Koponen, Aziz, Ramos, & Specia, 2012; Popović et al., 2014; Temnikova, 2010). Most studies involving PE effort have focused on languages with relatively sparse morphology. As pointed out by both Popović et al. (2014, p. 197) and Koponen (2016, pp. 51–52), research is needed to investigate to what extent these prior findings and assumptions about the PE process, edit types and effort are applicable to languages with different structures and richer morphology.

However, relatively few studies have investigated the correctness of the post-editor corrections, often working on the assumption that the edits were both correct and necessary to improve the MT text. This assumption is not entirely unproblematic, as studies have shown that errors may still remain after post-editing or even be introduced by the post-editors (de Almeida, 2013; Koponen & Salmi, 2015). Another question is whether all the corrections were necessary, particularly in cases of so-called “light PE” (ISO/DIS 18587, 2016), where stylistic editing is generally not needed. Post-editors may also have strong preferences for specific wording (Koponen, 2013) or stylistic preferences (Flanagan & Christensen, 2014) which lead them to make unnecessary changes. In the study by de Almeida (2013, pp. 189, 192), up to 25% of the edits analysed were classified as preferential rather than essential.

To deal with these open questions, this article presents a study examining PE quality from the perspective of the correctness and necessity of PE corrections. Based on an analysis of an English–

Finnish MT post-edited by translation students, we investigate the correctness and necessity of the edits. Correctness is evaluated in terms of accuracy of meaning as well as the grammaticality of the target language (TL). Necessity is defined based on whether the edits were essential to correct the meaning or language or whether they appear to be preferential edits related to style or word choices. The remainder of the article is arranged as follows: Section 2 presents background on both the Finnish context and PE research. Section 3 details the material and methods used in the study and Section 4 presents the results. The findings and their potential implications for PE practice and post-editor training are discussed in Section 5, and Section 6 presents concluding remarks and future work.

## 2. Background: MT and PE in the Finnish context

The use and usability of MT and PE vary greatly for different language pairs, depending on the quality of the MT available. Finnish is one language that has proved to be difficult for MT systems. This can be seen in MT evaluation campaigns, where even the best results for MT to or from Finnish lag behind most European language pairs (e.g., Bojar et al., 2016, p. 141). The MT quality issues, in turn, affect the extent to which MT is usable for PE. In the European Commission trials, English–Finnish MT was found sufficient to suggest ideas for expressions, at best, or even not usable at all (Leal Fontes, 2013, p. 11). In a survey of 238 Finnish translators, Mikhailov (2015) found that they did not consider MT important in the field, owing to the poor quality of the MT systems available.

MT quality issues are often linked to the specific characteristics of the Finnish language, which include rich inflectional morphology – each noun can have approximately 2,000 forms and each verb more than 12,000 forms – and long compound words written together without spaces or hyphens (Koskenniemi et al., 2012, p. 47). Furthermore, in Finnish, word order is relatively free, with syntactic and semantic relations between words expressed by the morphological form rather than the word order. In addition to MT quality, we hypothesize that these characteristics may also affect PE and the effort involved in the process.

### 2.1. The effect of MT errors on PE

Recent studies combining translation process research methodologies and MT error analyses have examined the effort involved in PE. As defined by Krings (1994/2001, p. 178), PE effort involves three dimensions: temporal, technical and cognitive effort. The approaches used to investigate these different aspects include: PE time and think-aloud protocols (Krings, 1994/2001), subjective effort assessment (Koponen, 2012), PE time and keylogging data (Koponen et al., 2012; Popović et al., 2014), pauses observed in keylogging (Lacruz et al., 2014; O’Brien, 2005) and keylogging combined with eye-tracking data (Daems et al., 2015).

In addition to process data, studies have examined the MT errors themselves, often through the changes effected by the post-editors. Temnikova (2010) has suggested a classification for MT errors ranked in terms of presumed cognitive effort. It contains ten error types: incorrect word forms, stylistically incorrect synonyms, incorrect words, extra words, missing words, mistranslated idioms, two types of punctuation error and two types of word-order error. A slightly modified version of this classification was used by Koponen et al. (2012), who also analysed whether an “incorrect word” involved a correct or an incorrect part of speech or an untranslated word. Lacruz et al. (2014, p. 78) defined five error types (mistranslation, omission or addition, syntax, word form and punctuation), also distinguishing between mechanical and transfer errors. A similar categorization was used by Popovíc et al. (2014). Categorizations involving edit, rather than error, types have also been used, for example, in Koponen (2012, p. 185), where edited words were labelled as changed form, changed word, changed part of speech or word-order changes such as inserted, deleted or moved. Most of the methods have examined MT errors or edits in terms of individual words, but Blain, Senellart, Schwenk, Plitt, & Roturier (2011) suggested combining edits into Post-Editing Actions (PEA) which consist of sets of interrelated edits. An example given by Blain et al. (2011, p. 165) involves changing a noun, which then also makes it necessary to make changes to its attributes – for example, to change adjective gender.

The results of these studies have provided increasing evidence that some types of MT error are easier to correct than others, and some types of edit performed by post-editors involve less effort than others. Generally, less effort appears to be connected to editing incorrect word forms, whereas word-order errors and reordering edits appear to involve greater effort (Daems et al., 2015; Koponen et al., 2012; Lacruz et al., 2014; Popovíc et al., 2014). Various studies have associated greater effort with part-of-speech errors (Koponen et al., 2012), mistranslated idioms and punctuation errors (Temnikova, 2010; Temnikova, Zaghouani, Vogel, & Habash, 2016), mistranslations and omissions or additions (Lacruz et al., 2014), mistranslated words (Popovíc et al., 2014), and mistranslations and structural errors (Daems et al., 2015). However, the association between any specific error and effort is not always straightforward. As Lacruz et al. (2014, p. 78) point out, the same error may be more or less severe depending on the context, and on whether the meaning (transfer error) or the language (mechanical error) is affected (see Koby & Champe, 2013, p. 166).

Relatively few studies so far have specifically addressed MT errors and PE in Finnish. Two studies involving the use of compound splitting and morphological segmentation for English–Finnish MT offer some insight into the common issues. Comparing MT to human translation, Tiedemann, Ginter and Kanerva (2015, p. 182) have shown that common inflectional forms such as the nominative case (nouns) and third person singular (verbs) are over-represented in Finnish MT, while other forms like genitives and passive or participial verbs are under-represented. Pirinen et al. (2016, pp. 63-65) also reported frequent errors related to morphological forms as well as problems with forming compounds in the MT. A Master's thesis by Nordman (2015) compared English-Finnish MT by three different MT engines and found that incorrect words accounted for 25.4% of the errors and missing relations between words for 26.6%. However, incorrect meanings were more frequent than incorrect relations with the rule-based engine, whereas statistical engines made more mistakes in incorrect relations (Nordman, 2015, pp. 22-23).

PE data involving Finnish as the target language has been discussed in two studies. Koponen (2013) examined post-editor choices by 11 participants when editing 139 short controlled-language MT sentences. Koponen and Salmi (2015) investigated a monolingual PE task where 48 participants corrected English-Finnish MT without access to the source text. That study mainly addressed correctness at the sentence level, but based on an analysis of specific cases that appeared particularly difficult to correct, these often involved missing or incorrect words and word-form errors that obscured the syntactic and semantic relations between words (Koponen & Salmi, 2015, p. 133). This last finding particularly suggests that whereas word-form errors have often been suggested to be among the easiest to correct, the situation may be different in a morphologically rich language such as Finnish.

## 2.2. The quality of post-editor corrections

Studies on MT errors and PE often rely on the assumption that the changes performed by the post-editors are correct and represent actual errors in the MT. As Koponen (2016, p. 51) argues, this assumption is not always correct. When the quality of PE corrections themselves has been analysed, errors needing correction may still appear, and sometimes post-editors even introduce new errors. This was observed in the study reported in Koponen and Salmi (2015), where the students acting as post-editors were not particularly attentive to language errors, as nearly half of the post-edited sentences contained language errors. Examples include unedited punctuation errors, typographical errors made by post-editors, or grammatical errors caused by changing part of the sentence but failing to edit some other part accordingly (Koponen & Salmi, 2015, p. 131).

A more detailed view of correctness is presented by de Almeida (2013), who analysed whether post-editors had performed all the essential changes, defined as changes needed to make the sentence grammatically correct and semantically accurate (2013, p. 100). The results show that the post-editors failed to make such essential changes in 11% to 15% of analysed cases (depending on language pair) and, moreover, introduced new errors in approximately 5% of cases (de Almeida, 2013, pp. 189–194). According to de Almeida (2013, pp. 190–191, 193–194), most errors in both categories involved mistranslations and language errors. Temizöz (2016) also analysed the types of error found in a PE task comparing professional translators and subject-matter experts. Most errors made by the translators involved terminology or language errors, whereas the experts made mostly language errors, with small numbers of mistranslations and accuracy and consistency errors in both groups (Temizöz, 2016, pp. 10–11).

Another question related to PE quality is whether all the corrections are, in fact, necessary. Some edits may involve changes that de Almeida (2013, p. 100) calls "preferential": situations where the sentence would be grammatically correct and accurate even without the change. According to de Almeida (2013), preferential changes accounted for 16% to 25% of the cases analysed (depending on language pair) and involved mainly language issues or lexical choices (2013, pp. 189–192). Changes deemed unnecessary may stem from different interpretations of what, precisely, must be changed or corrected. Although PE guidelines have been drafted (e.g., ISO/DIS 18587, 2016; TAUS, 2010), their practical implementation is not necessarily clear to post-editors. Flanagan and Christensen (2014) studied three translator trainees who used the PE guidelines by TAUS (2010) in a PE task. Their analysis of retrospective interviews, reflective essays and the PE texts showed that while guidelines concerning grammatical, syntactic and semantic correctness appeared to be clear, the trainees had differing interpretations of the guidelines regarding style and, in retrospect, reported spending much time on unnecessary stylistic changes (Flanagan & Christensen, 2014, pp. 264–265).

Different post-editors may also correct the same error in different ways. A comparative study of different PE versions of Finnish MT sentences conducted by Koponen (2013) indicated that most variation appears in idiomatic or ambiguous expressions. Such variation may indicate increased cognitive processing. Passages where all the translators or post-editors make the same choice have been argued to be cognitively easier than those where multiple differing versions appear (Campbell, 2000; Englund Dimitrova, 2005; Krings, 1994/2001). This is supported by O'Brien (2005, p. 55), who found pauses in keylogging data (indicating increased cognitive processing) in passages where post-editors produced multiple versions.

## 3. Correctness and necessity of PE corrections: study set-up

### 3.1. Material and data collection

The data analysed in this study form a subset of English–Finnish PE data collected in October 2016 in an experiment where 16 translation students post-edited a short MT text as part of a course on MT and PE at the University of Helsinki, Finland. All the participants in the experiment were native Finnish speakers studying translation in different language programmes (English, German or Russian). The PE task was carried out using Translog-II (Carl, 2012). The dataset also includes keylogging and eyetracking data, but only the MT and PE texts produced by the participants are used in this study, as our focus is on the edits.

The purpose of this pilot stage of the study was to explore suitable approaches and categories for a larger-scale analysis of multiple differing PE corrections and their relationship to MT errors. To provide for a more detailed investigation, we limited the scope of this pilot stage to a smaller subset of PE versions. For this purpose, we decided to analyse PE versions from five translation students only. These five were selected because they indicated that English–Finnish was their main language pair and therefore they could be expected to have high proficiency in the source language (SL). Of the five participants whose data were used in this study, four are Master's students and one is in the final stages of a Bachelor's degree. Two had had no translation experience outside their studies, two indicated that they had done some translation assignments but not regularly, and one had worked as a translator for three months. All the students had a theoretical knowledge of and practical experience with PE during their studies, but no professional experience.

The English ST was obtained from the WMT16 test set for the English–Finnish news translation task (Bojar et al., 2016). The ST, originally from the BBC website, contains 27 segments

(385 words) and provides instructions on how readers can send material to the BBC. The text was selected because it was of a suitable length to be edited in one data-collection session and it contained no specialized terminology. The MT output that was post-edited contains sentences from three different MT systems (rule-based, statistical and neural MT) – nine sentences from each. The same MT version was edited by all five participants. The participants were able to see the entire text at the same time and edit the text in any order they chose.

The participants were told to conduct light PE, which had been discussed in detail during the course before the experiment. The following instructions, based on the principles of light PE presented in ISO/DIS 18587 (2016), were given at the start of the task:

- Make use of the raw MT as much as possible.
- Aim to produce a translation that conveys the correct meaning and is grammatically correct.
- Check that there is no extra or missing information.
- Change sentence structure only if the meaning is incorrect or unclear.
- Follow Finnish spelling and punctuation conventions.

Only the information available in the ST and MT was to be used in the task; no external sources of information such as dictionaries were available. This decision was related to limitations set by the eyetracking equipment used in the data collection. No time limit was set for the task, because we wanted to allow the participants to work on the text until they considered it satisfactory; but they were advised to avoid spending excessive time on any one correction. The time taken to complete the task varied from 13 to 23 minutes.

### 3.2. Identifying PE changes

In order to identify the changes the participants made, we used the edit distance metric HTER (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006), which compares the MT and PE versions of a sentence and computes the minimum number of word-level changes divided by the number of words in the PE version. HTER also provides information about the edit type, classifying edits from the point of view of the MT as deletions (words appearing in the PE but not in the MT), insertions (words appearing in the MT but not in the PE), word substitutions or word-order shifts. For some languages, HTER has resources to recognize substitutions involving the same word stem and synonym substitutions, but this support is not available for Finnish. For a more fine-grained analysis of which substitutions involved changes to the morphological form of a word, we ran HTER also on lemmatized texts where the surface form of the word had been replaced with the base form (lemma) using the open-source morphological analysis tools OMorFi (Pirinen, 2008) and FinnPos (Silfverberg, Ruokolainen, Lindén, & Kurimo, 2016). The lemmatized versions were checked manually to ensure accuracy and potential errors were corrected.

The HTER data showing the MT words aligned with the corresponding PE words and the edit type were then used as the basis for a manual analysis, where each word was annotated with one of the following categories according to the actions in PE:

- unedited: no change;
- form changed: different morphological form;
- word changed: different lemma;
- deleted: word removed;
- inserted: word added;
- order: position of a word changed.

The reason for defining the analysis in terms of PE actions or edits rather than MT errors is that some edits may not indicate actual MT errors, but may instead represent preferential choices (see de Almeida, 2013, p. 100). Some edits were found to involve a combination of categories: for example, both word form and word order. Such cases were annotated with all the relevant categories in the manual analysis. Following the usual practice for edit distance metrics such as HTER, punctuation marks were treated as words for the purposes of this analysis. In deciding which words should be labelled as moved, if the order of parts of a sentence had been changed, the HTER alignment data were used, even though they do not necessarily match the actual edit operation.

### 3.3. Analysing the correctness and necessity of the edits

In the manual analysis, each word was annotated both for correctness and the necessity of the edit. *Correctness* was defined as accurate in terms of both ST meaning and proper grammar and spelling of the TL form. *Necessity* meant that the change was needed to make the TL sentence comprehensible, grammatically correct or accurate in meaning. Both authors first annotated the edits of participant A and discussed differing annotations in order to agree on how edits were categorized (see example in the next paragraph). Edits by the other participants were marked first by one author and then reviewed by the other, and the authors discussed cases where the correctness or necessity of a specific edit was unclear. These generally involved decisions about which lexical choices could be considered to reflect the ST meaning accurately. In the case of necessity, it was usually a question of deciding whether the word used by the participant was acceptable (correctness) or whether the word used in the MT was acceptable (necessity).

Some words contained multiple edits as described in 3.2 above. In such cases, each edit type was analysed separately for correctness and necessity. For example, if a form change was necessary, but a word-order change was not, this was marked as “mixed”. As the students had been instructed to do a light PE, we decided not to consider the cohesion or coherence of the text when evaluating TL correctness. The MT sometimes had two differing translations of the same source word: for example, *output* was translated in different sentences as *tuotanto* “production” and *julkaisu* “publication”. If both were semantically correct, changing one to ensure consistency was deemed not necessary in this case. This decision was made because the ST was intended to provide general information for a non-specialist readership rather than a specialized text where terminological consistency would be paramount.

## 4. Results

### 4.1. Types of edit identified in the PE versions

In total, 1,715 words were analysed in the PE versions. Table 1 shows a comparison of the amount of editing performed by each of the five participants, who are referred to using the letters A–E (see first column). The second column shows the total number of words in the lemmatized PE version. Compound words which are normally written as one word are treated as separate words here. The third column shows the number of unedited words and the fourth column shows the number of words edited. The last column shows the HTER score, which expresses the amount of editing calculated as the number of edited words divided by the total number of words in the PE version. In Table 1, we see that participant B has edited the largest number of words, participants A, C and D generally speaking edited a similar number of words, and participant E edited the text considerably less than the others.

Table 1. Number of edited and unedited words per participant

| Participant | Words | Unedited | Edited | HTER |
|---|---|---|---|---|
| A | 348 | 203 | 145 | 41.7 |
| B | 341 | 172 | 169 | 49.6 |
| C | 347 | 202 | 145 | 41.8 |
| D | 341 | 210 | 131 | 38.4 |
| E | 338 | 249 | 89 | 26.3 |

The number and percentage of edits classified in each category are shown in Table 2. Overall, 60.4% of the words were left unedited by the participants. The most common types of edit involve form changes and insertions, each of which accounts for nearly 10% of the edits. Order changes (2.9%) and cases involving multiple changes (2.3%) are less common than the other types. The cases with multiple changes were commonly word-form changes combined with order changes (27 out of 39 cases). As mentioned in 3.2 above, punctuation marks (238 cases) are counted as words in Tables 1 and 2. Overall, punctuation was mostly left unedited (200 cases), but there were 25 cases of deletion, 12 cases of insertion, and one case of substitution with a different punctuation mark.

Table 2. Categorization of edits

| Edited or unedited words | Number | % |
|---|---|---|
| Unedited words | 1,036 | 60.4 |
| Edited words | 679 | 39.6 |
| – Form changed | 165 | 9.6 |
| – Word changed | 116 | 6.8 |
| – Deleted | 148 | 8.6 |
| – Inserted | 162 | 9.4 |
| – Order changed | 49 | 2.9 |
| – Multiple changes | 39 | 2.3 |
| Total words | 1,715 | 100 |

### 4.2. Correctness and necessity of edits

Nearly all of the unedited words were deemed correct, but in 36 cases (3% of unedited words) a necessary correction had not been made, and the PE was in fact incorrect. The majority (24 cases) involved language errors which do not necessarily affect meaning, but four involve cases where missing words have not been added, leading to missing information. The results for correctness and necessity of edits are shown in Table 3.

Table 3. Correctness and necessity of edited words

| | Necessary | Unnecessary | Mixed | Total |
|---|---|---|---|---|
| Correct | 377 | 235 | 8 | 620 |
| Incorrect | 20 | 36 | 3 | 59 |
| Total | 397 | 271 | 11 | 679 |

Nearly all (620, 91%) of the 679 edits were evaluated as correct in terms of meaning and grammar. Of those 620 cases, 377 (61%) were also considered necessary, since these cases represent errors in the MT. In contrast, 235 cases (38%) were deemed unnecessary, meaning that the MT sentence was already correct without the change. In eight cases (1%) labelled as “mixed”, multiple changes had been performed but only some of them were necessary. Most of these (five) were combinations of word-form and word-order changes, where only the word-form change was necessary. The 59 cases (9%) labelled as “incorrect” represent situations where the PE version contains an error. Of these, 20 (34%) are cases where a necessary correction was attempted, but the edit did not succeed in correcting the error. These represent cases where the participant changed an incorrect word or word form, but the substitution contained some error either due to wrong word or form choice, or typographical errors. In 36 cases (61%), the incorrect word resulted from an unnecessary change, and in three cases (5%) the change was only partially necessary. These represent instances where a new error was introduced by the participant, and most cases (15) involved unnecessary deletions.

## 5. Discussion

The categorization of the edit types in the previous section shows that nearly 25% of all the edits performed were changes to word forms. This finding is in line with Tiedemann et al. (2015) and Pirinen et al. (2016), who also noted frequent errors related to morphological forms in Finnish MT. A closer analysis of the edit types also reveals that the majority of word-form edits (70%) were necessary, as were most insertions (84%) and word substitutions (67%). These three edit types are both frequent and mostly necessary, which indicates that they represent frequent errors in Finnish MT. In contrast, 80% of word-order changes and 70% of deletions were deemed unnecessary. Word-order changes were also infrequent and the fact that most of them were unnecessary can be explained by the fact that Finnish has a relatively free word order. Most word-order changes appear to have been stylistic. On the other hand, deletions were relatively frequent but also mostly unnecessary, and 10% of them also resulted in incorrectly omitting information. Most deletions, however, involved personal pronoun subjects, which are optional in Finnish. Changes to punctuation were also generally deemed unnecessary.

The number of words edited varied between the participants, as can be seen in their HTER scores (Table 1). This finding is similar to that in previous studies. Our analysis also showed a variation in correctness and necessity corrections among the participants. Participant B edited the highest number of words (169), and also had the highest number of incorrect and unnecessary edits (17, or 10%). Participant E, on the other hand, edited the lowest number of words (89) and had the highest proportion of correct and necessary edits (64%), but also the second-highest number of incorrect and unnecessary edits (7, or 8%). More than half (20 out of 35, or 57%) of the instances where a necessary edit had not been performed were found in the version by participant E. Previous studies have also observed variation in edits, which appears to be partly due to personal preferences. Preferential edits were common in the study by de Almeida (2013), and Koponen (2013) also found that preferences for specific wording can influence PE choices quite strongly.

Different post-editors, especially those less used to PE, also appear to interpret the task differently, particularly in terms of what they think should be changed. In our study, this can be seen in the number of unnecessary changes. A parallel can be drawn with Flanagan and Christensen (2014, pp. 264–265), who discussed the different ways their three students interpreted PE guidelines about accuracy and style, and the students' own observations about spending too much time on stylistic corrections. This difficulty in interpreting PE guidelines could be mitigated with more detailed information and training regarding necessary and unnecessary edits, as Flanagan and Christensen (2014) also note. Our results suggest that specific edit types merit closer attention, at least for English–Finnish PE. In particular, word-order changes appear to be largely stylistic, and therefore unnecessary, at least in the case of a light PE. Deletions also appear to be mostly unnecessary, and may even lead to incorrectly omitting information. Both insertions and word and form changes, on the other hand, were generally found to be necessary.

Our approach to correctness is similar to that of Koponen and Salmi (2015), but we evaluated correctness on the word rather than sentence level. Unlike in Koponen and Salmi (2015), however, we combined meaning and language in the evaluation of correctness. This had the disadvantage that semantic, grammatical and typographical errors were included in the same category, although their effect on the MT or PE text may be different. This is related to the distinction between mechanical and transfer errors (Koby & Champe, 2013; Lacruz et al., 2014), and a closer analysis of the edits would be needed to determine how different types of edit affect the meaning or language.

Determining the necessity of edits was not always straightforward. An example of this was the use of the second person, which has only one form in English. In contrast, Finnish has distinct forms for singular and plural, and the plural is also used as a politeness form. Because the ST contained mostly instructions for how to share content, second-person forms were frequent, and a mixture of plural and singular was used in the MT sentences. As explained in 3.3 above, edits were deemed unnecessary if an otherwise correct verb form was changed from plural to singular or vice versa, to make the usage consistent across sentences. This decision was made due to the task description involving light PE, but in a full PE task consistency would be more relevant. The word-level approach was also sometimes difficult. As Blain et al. (2011) observe, some edits are interrelated: changing one word may result in other changes becoming necessary. In Finnish, substituting one verb with another, for example, may lead to changing the case ending of nouns governed by the verb. In this study, the changes resulting from other edits were labelled as unnecessary if some other option, such as selecting a synonymous verb, would have been possible without editing other parts of the sentence.

Based on the PE versions alone, it is also not possible to determine the amount of effort involved in each edit. An analysis of the process data would give more insight, but was outside of the scope of this article. However, we suggest that the edit with the smallest number of changes is not necessarily always the easiest in terms of cognitive effort. This is demonstrated in the following examples. Example (1) shows the ST passage and the MT version:

(1) We will publish your name as you provide it (unless you ask us not to) but we will never publish your mobile phone number.
*Julkaisemme nimenne niin kuin annatte sen (paitsi jos pyydätte meitä olemaan), mutta emme koskaan julkaise kännykkänumeroanne. (machine translation, our emphasis)*
= [We publish your name as you (pl.) provide it (except if you (pl.) ask us to be) but we never publish your (pl.) mobile number.]

An elliptical structure directly corresponding to the English *unless you ask us not to* is not possible in Finnish, and the MT is incorrect, changing the meaning completely. Examples 2–6 show the five different PE versions of the relevant underlined passage in Example 1.

(2) unless you ask us not to
*paitsi jos pyydät meitä olemaan julkaisematta (Participant C)*
= [except if you (sg.) ask us not to publish]

(3) unless you ask us not to
*paitsi jos pyydätte meitä käyttämättä (Participant E)*
= [except if you (pl.) ask us without using]

(4) unless you ask us not to
*paitsi jos pyydätte ettemme käytä sitä (Participant A)*
= [except if you (pl.) ask that we not use it]

(5) unless you ask us not to
*paitsi jos kiellätte sen (Participant D)*
= [except if you (pl.) forbid it]

(6) unless you ask us not to
*paitsi jos erikseen kiellät sen (Participant B)*
= [except if you (sg.) specifically forbid it]

Perhaps the most obvious solution is shown in Example 2, where C added the word *julkaisematta* (literally: "be without publishing"). Similar approaches were taken by participants A and E, but E deleted the word *olemaan*, making the PE version ungrammatical (Example 3), while A also changed the structure of the sentence (Example 4). Participants D and B, however, replaced the verb *pyytää* "ask, request" with *kieltää* "deny, forbid" and added the word *sen* "it", changing the sentence structure even further. While the choice made by participants A, C and E may appear more straightforward than the structural change by D and B, it is difficult to determine which involved more effort. Using the structure *olemaan julkaisematta* is somewhat unusual, although correct, in Finnish, while the option of using the word *kieltää* is more idiomatic. A closer analysis of effort indicators in the process data would be needed to determine the effort involved in each approach.

It should also be noted that the participants in this pilot study were translation students, and a study involving professional translators might have produced different results. The participants did, however, have some experience with translation and training in PE, so they can be considered semiprofessionals (Englund Dimitrova, 2005, p. 16). Further studies involving professional translators and longer texts would be necessary.

## 6. Conclusion and future work

In this article we have presented a pilot study aiming to investigate PE quality in a light PE task. Quality was investigated in terms of two dimensions: the correctness and necessity of the edits. Edits made by five translation students during a light PE task involving English–Finnish MT were categorized in terms of these two dimensions. The results show that while most edits were correct, 38% of them were in fact unnecessary to correct meaning and language. Further analysis also suggests that for this language pair, specific edit types – word-order changes and deletions – are largely unnecessary, which may have implications for PE practice and training. The MT version post-edited by the participants contained translations from three different types of MT engine (rule-based, statistical and neural). While a detailed comparison of these different MT types was not within the scope of this article, future work will also involve a comparative analysis of errors by different types of MT engine.

Since the purpose of this study was to function as a pilot to explore approaches to analysing PE corrections by different editors, it was found necessary to limit the number of PE versions analysed. The small sample size naturally limits the generalizability of the results. Having the correctness and necessity of edits assessed by the two authors only is to some extent also subjective. However, the findings of this pilot study provide useful starting points, and this stage will be followed by a more extensive analysis of PE data from a larger number of participants. In future work we also plan to have the assessment of correctness and necessity conducted by other independent evaluators for increased objectivity.

The use of lemmatization tools was found useful in the analysis, and they revealed that a large proportion of substitutions in the PE versions in fact involved word forms. The majority of word-form changes were also found to be necessary, which supports earlier findings that word-form errors are frequent in Finnish MT. To determine the effect of word-form errors on the post-editor and the effort involved in correction, a more detailed analysis of the PE process would be needed. Future work will include the use of keylogging and eyetracking data to analyse the effort indicators related to performing different types of edit. Such analyses could reveal more about the relative cognitive effort connected to specific errors and error types.

## References

Blain, F., Senellart, J., Schwenk, H., Plitt, M., & Roturier J. (2011). Qualitative analysis of post-editing for high quality machine translation. In *Proceedings of the 13th Machine Translation Summit* (pp. 164–171). Asia-Pacific Association for Machine Translation.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y. Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., & Zampieri, M. (2016). Findings of the 2016 Conference on machine translation. In *Proceedings of the First Conference on Machine Translation (WMT)* (pp. 131–198). Stroudsburg, PA: The Association for Computational Linguistics.

Campbell, S. (2000). Critical structures in the evaluation of translations from Arabic into English as a second language. *The Translator, 6*, 37–58.

Carl, M. (2012). Translog – II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* (pp. 4108–4112). European Language Resources Association.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2015). The impact of machine translation error types on post-editing effort indicators. In S. O'Brien & M. Simard (Eds.), *Proceedings of Fourth Workshop on Post-editing Technology and Practice (WPTP4)* (pp. 31–45). Association for Machine Translation in the Americas.

de Almeida, G. (2013). *Translating the post-editor: An investigation of post-editing changes and correlations with professional experience across two romance languages* (Doctoral thesis, Dublin City University, Dublin, Ireland). Retrieved from http://doras.dcu.ie/17732/

Englund Dimitrova, B. (2005). *Expertise and explicitation in the translation process*. Amsterdam: John Benjamins.

Flanagan, M., & Christensen, T.P. (2014). Testing post-editing guidelines: How translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer, 8*(2), 276–294.

ISO/DIS 18587. (2016). Translation Services – Post-editing of machine translation output: Requirements. Geneva: International Organization for Standardization.

Koby, G.S., & Champe, G.G. (2013). Welcome to the real world: Professional-level translator certification. *Translation & Interpreting, 5*(1), 156–173.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *7th Workshop on Statistical Machine Translation. Proceedings of the Workshop* (pp. 181–190). Stroudsburg, PA: Association for Computational Linguistics.

Koponen, M. (2013). This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In S. O'Brien, M. Simard & L. Specia (Eds.), *Workshop Proceeding: Workshop on Post-editing Technology and Practice (WPTP-2)* (pp. 1–9). Allschwil: The European Association for Machine Translation.

Koponen, M. (2016). *Machine translation post-editing and effort: Empirical studies on the post-editing process* (Doctoral thesis, University of Helsinki, Helsinki, Finland). Retrieved from http://hdl.handle.net/10138/160256

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In S. O'Brien, M. Simard & L. Specia (Eds.), *Proceedings of WPTP 2012. AMTA 2012 Workshop on Post-editing Technology and Practice* (pp. 11–20).

Koponen, M., & Salmi, L. (2015). On the correctness of machine translation: A machine translation post-editing task. *Journal of Specialised Translation, 23*, 118–136.

Koskenniemi, K., Lindén, K., Carlson, L., Vainio, M., Arppe, A., Lennes, M., Westerlund, H., Hyvärinen, M., Bartis, I., Nuolijärvi, P., & Piehl, A. (2012). *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. META-NET White Paper Series. Heidelberg: Springer.

Krings, H.P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing process*. G. S. Koby (Translated into English. Original *Texte Reparieren*, 1994). Kent, OH: The Kent State University Press.

Lacruz, I., Denkowski, M., & Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. In S. O'Brien, M. Simard & L. Specia (Eds.), *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)* (pp. 73–84). Association for Machine Translation in the Americas.

Leal Fontes, H. (2013). Evaluating machine translation: Preliminary findings from the first DGT-wide translators' survey. *Languages and Translation, 6*, 10–11.

Mikhailov, M. (2015). Minor language, major challenges: The results of a survey into the IT competences of Finnish translators. *The Journal of Specialised Translation, 24*, 952–975.

Nordman, M. (2015). *"Aina irrota vedenkeitin jahka ei kotona apu": käyttöohjeiden konekäännösten virheanalyysi. ["Always unplug kettle until no help at home": error analysis of machine translations of kettle user instructions.]* (Unpublished master's thesis). University of Turku, Turku, Finland.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation, 19*(1), 37–58.

Pirinen. T. (2008). *Automatic finite state morphological analysis of Finnish language using open source resources (in Finnish)* (Unpublished master's thesis). University of Helsinki, Helsinki, Finland.

Pirinen, T., Toral, A., & Rubino, R. (2016). Rule-based and statistical morph segments in English-Finnish SMT. In T.A. Pirinen, E. Simon, F.M. Tyers & V. Vincze (Eds.), *Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages* (pp. 56–69). Retrieved from http://rgai.inf.u-szeged.hu/project/iwclul/proceedings.pdf

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics, 93*, 7–16.

Popović, M., Lommel, A., Burchardt, A., Avramidis, E.& Uszkoreit, H. (2014). Relations between different types of post-editing operations, cognitive effort and temporal effort. In *EAMT 2014: 17th Annual Conference of the European Association for Machine Translation* (pp. 191–198). Allschwil: The European Association for Machine Translation.

Silfverberg, M., Ruokolainen, T., Lindén, K. and Kurimo, M. (2016). FinnPos: An open-source morphological tagging and lemmatization toolkit for Finnish. *Language Resources and Evaluation, 50*(4), 863–878.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation* (pp. 223–231). Association for Machine Translation of the Americas.

TAUS. (2010). *Machine translation post-editing guidelines*. Retrieved from https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines.

Temizöz, Ö. (2016). Postediting machine translation output: Subject-matter experts versus professional translators. *Perspectives, 24*(4), 646–665.

Temnikova, I. (2010). A cognitive evaluation approach for a controlled language post-editing experiment. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluati*on *(LREC'10).* European Language Resources Association (ELRA).

Temnikova, I., Zaghouani, W., Vogel, S., & Habash. N. (2016). Applying the cognitive machine translation evaluation approach to Arabic. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 3644–3651). European Language Resources Association (ELRA).

Tiedemann, J., Ginter, F., & Kanerva, J. (2015). Morphological segmentation and OPUS for Finnish–English machine translation. In *Tenth Workshop on Statistical Machine Translation. Proceedings of the Workshop* (pp. 177–183). Red Hook, NY: Association for Computational Linguistics.