Machine Translation
https://doi.org/10.1007/s10590-019-09228-7

A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output

Maarit Koponen[1] · **Leena Salmi**[1] · **Markku Nikulin**[1]

Received: 15 July 2018 / Accepted: 9 February 2019

Abstract
This paper presents a comparison of post-editing (PE) changes performed on English-to-Finnish neural (NMT), rule-based (RBMT) and statistical machine translation (SMT) output, combining a product-based and a process-based approach. A total of 33 translation students acted as participants in a PE experiment providing both post-edited texts and edit process data. Our product-based analysis of the post-edited texts shows statistically significant differences in the distribution of edit types between machine translation systems. Deletions were the most common edit type for the RBMT, insertions for the SMT, and word form changes as well as word substitutions for the NMT system. The results also show significant differences in the correctness and necessity of the edits, particularly in the form of a large number of unnecessary edits in the RBMT output. Problems related to certain verb forms and ambiguity were observed for NMT and SMT, while RBMT was more likely to handle them correctly. Process-based comparison of effort indicators shows a slight increase of keystrokes per word for NMT output, and a slight decrease in average pause length for NMT compared to RBMT and SMT in specific text blocks. A statistically significant difference was observed in the number of visits per sub-segment, which is lower for NMT than for RBMT and SMT. The results suggest that although different types of edits were needed to outputs from NMT, RBMT and SMT systems, the difference is not necessarily reflected in process-based effort indicators.

Keywords Neural machine translation · Statistical machine translation · Rule-based machine translation · Post-editing · Translation process research · Keylogging

✉ Maarit Koponen
maarit.koponen@utu.fi

[1] School of Languages and Translation Studies, University of Turku, 20014 Turku, Finland

Published online: 08 March 2019

1 Introduction

Recent developments in neural machine translation (NMT) and reported quality improvements over phrase-based statistical machine translation (SMT) have led to much excitement. NMT systems have outperformed other types in recent studies and evaluation campaigns in many language pairs. The situation, however, varies in different language pairs: languages with highly productive morphology, for example, have overall been found to be challenging for MT systems. Recent error analyses comparing NMT to SMT systems, however, suggest that NMT produces more fluent output in morphologically rich languages (Toral and Sánchez-Cartagena 2017; Klubička et al. 2017, 2018). One issue affecting quality in many languages is posed by the fact that the necessary language resources are not equally available for all languages or language pairs. "Low-resource settings" are identified by Koehn and Knowles (2017, p. 28) as one of the key challenges for NMT limiting quality improvements. While some recent studies comparing NMT and SMT systems suggest that NMT produces fewer word form errors, offering potential improvements for morphologically rich languages, the comparative lack of resources still poses issues in the case of Finnish.

The use of MT combined with post-editing (PE) has become increasingly integrated in workflows in the translation industry (e.g. Plitt and Masselot 2010) as well as in translator education, with the most recent development being interactive MT systems (e.g. Green et al. 2014; Peris and Casacuberta 2018). The uptake of MT and PE varies in different countries and language pairs, as well, generally due to differences in the output quality of available MT systems, as the use of low-quality MT as a raw version to be post-edited is not feasible. When assessing the usefulness of MT for practical purposes like PE, the effort involved in correcting the MT errors is a central issue, which has received increased research interest in recent years. Although recent studies suggest that NMT can reduce different types of errors and produce more fluent output also in morphologically rich languages, the effect of NMT on PE effort has so far been studied little. Bentivogli et al. (2016, 2018) reported reduced PE effort for NMT versus SMT, however, effort was assessed only based on the number of changes identified in the final PE product using the HTER metric. As pointed out by Daems (2016, p. 32), a product-based analysis of the changes evident in the final post-edited version alone cannot measure the actual effort involved. Prior research focusing on PE effort (Koponen 2012; Popović et al. 2014; Lacruz et al. 2014) has also demonstrated that while the number of errors is one factor affecting the PE effort needed, the type of errors is another major factor: some errors are more demanding than others. So far, only a few studies have examined PE effort in NMT versus SMT or rule-based MT (RBMT) using process measures such as keystrokes or pauses. Reductions in temporal and technical effort have been reported for some language pairs (Castilho et al. 2017; Toral et al. 2018). In the case of many languages, for example Finnish, the relationship between different types of MT errors and PE effort remains largely unresearched to date, and more studies are needed to explore the issue.

To address the question of MT quality and PE effort, this paper presents an analysis of PE product and process data collected from a total of 33 student post-editors during a task of post-editing a short machine translated from English to Finnish using three different MT systems (neural, rule-based and statistical). The objective of the study is to determine potential differences (1) in types of edits performed on output from each system, and (2) in effort indicators identified in PE process data, especially comparing NMT to RBMT and to SMT. The product-based analysis follows the edit type analysis and PE quality assessment developed for the pilot reported in Koponen and Salmi (2017). We analyze the types of edits performed by the participants (word substitutions, word form changes, word order changes, additions and deletions), and compare the types across different system outputs. The comparison of edit types is complemented by an analysis of the correctness and necessity of the edits to determine to which extent the edits reflect actual MT errors. In addition, we use a process-based approach to analyze PE effort metrics identified in the keylogging data. To measure technical PE effort, we examine the number of times a specific passage is edited and the number of keystrokes used. To measure cognitive and temporal effort, we examine the number and length of pauses occurring during the editing of a specific passage. Section 2 provides an overview of the theoretical background and prior research into NMT quality as well as effort in the PE process and PE quality. The data and methods used in this study are described in Sect. 3. Section 4 presents the results of the product and process-based analyses, as well as examples of certain characteristic edit types in NMT, RBMT and SMT, which are further discussed in Sect. 5. Finally, conclusions and implications of this study are presented in Sect. 6.

2 Background and related work

2.1 NMT output and quality

The use of NMT has rapidly increased in a short period of time. Several studies have shown that it outperforms other types of systems in terms of automatic scores and human evaluations. Different types of analyses have been used to compare the quality of NMT and SMT (and less commonly RBMT). Some studies have used automatic metrics like BLEU (Papineni et al. 2002) or HTER (Snover et al. 2006) comparing MT output to "gold standard" human translations or post-edited versions, and human evaluations of quality, often in terms of fluency and adequacy ratings (e.g. Wu et al. 2016; Junczys-Dowmunt et al. 2016; Crego et al. 2016; Castilho et al. 2017). However, recent research suggests that automatic metrics may not always be suitable: Shterionov et al. (2018) compared automatic evaluation scores to human evaluation and noticed that the automatic scores underestimated the quality of NMT systems. Other studies, like Castilho et al. (2017), report mixed results using different automatic (HTER, BLEU) and human evaluation metrics (fluency, adequacy) – SMT systems outperformed NMT in two case studies out of three – and point out that results vary depending on domain and language pair.

More detailed error analyses of NMT in various language pairs have also been conducted using (semi-)automatic methods (Toral and Sánchez-Cartagena 2017; Bentivogli et al. 2018) or manual error analyses (Klubička et al. 2017, 2018; Popović 2018) or a combination (Burchardt et al. 2017). Bentivogli et al. (2018) found that in English-to-German and English-to-French, the NMT system reduced errors considerably particularly for morphology and overall word order, but its weaknesses were in lexical choice and reordering of specific cases where semantic understanding is needed. Popović's (2018) study covered the language pairs German-to-English, English-to-German and English-to-Serbian, and found that NMT was better than a phrase-based SMT system in generating verb forms, avoiding verb omissions and handling English noun collocations and German compound words, while prepositions and (English) ambiguous words caused problems to NMT. Burchardt et al. (2017) compared NMT to both SMT and RBMT in German-to-English and English-to-German, and observed that NMT performs best on various features like coordination and ellipsis, multi-word expressions, long distance dependencies and named entities, while RBMT outperforms both NMT and SMT for verb tense, aspect and mood, as well as ambiguity. The study by Toral and Sánchez-Cartagena (2017) covered six languages from four different families (Germanic, Slavic, Romance and Finno-Ugric), and found that the best NMT system clearly outperformed the best phrase-based SMT system for all language directions out of English. While their results showed that translations produced by NMT systems were more fluent, and more accurate in terms of word order and inflected forms, NMT performed poorly when translating very long sentences (Toral and Sánchez-Cartagena 2017). These recent error analyses suggest that NMT reduces different types of errors and produces more fluent output specifically in morphologically rich languages like Czech, Finnish (Toral and Sánchez-Cartagena 2017), Serbian (Popović 2018) and Croatian (Klubička et al. 2017, 2018).

Morphologically rich languages like Finnish have overall been found to be challenging for MT systems. Results for (SMT) systems translating to or from Finnish have generally lagged behind most large European languages, in terms of BLEU scores as well as human evaluations (e.g. Bojar et al. 2016; Leal Fontes 2013). Due to the quality issues, MT to or from Finnish has not been widely used in commercial and professional contexts, although SMT (e.g. Tiedemann et al. 2015; Pirinen et al. 2016), NMT (e.g. Östling et al. 2017; Grönroos et al. 2017) and RBMT (Hurskainen and Tiedemann 2017) systems have been developed in academic settings. Finnish is included in online systems like Google Translate, and some commercial RBMT systems exist for the language pairs English-Finnish-English (Sunda) and Finnish-English (TranSmart), and reports from the field also indicate integration of proprietary (S)MT systems by some translation service providers like Lingsoft (see Ervasti 2017). Finnish is also included in the European Commission eTranslation system (European Commission 2018), and was one of the first languages for which the new NMT system was implemented in autumn 2017. Quality issues of MT to/from Finnish have been linked to specific characteristics of the Finnish language, including rich inflectional morphology, productive derivation and long compound words, as well as relatively free word order (Koskenniemi et al. 2012, p. 47; Tiedemann et al. 2015; Grönroos et al. 2017). Prior studies involving SMT quality have shown that

problems related to morphological errors and errors in forming compound words are common (Tiedemann et al. 2015; Pirinen et al. 2016). The only previous study examining NMT quality with Finnish as a target language by Toral and Sánchez-Cartagena (2017) reported that NMT reduced inflection errors by 11.65% and word order errors by 12.12%, suggesting that NMT offers potential improvements for Finnish MT.

2.2 Machine translation errors and post-editing effort: product and process

As defined in Krings' (2001) seminal work on post-editing and effort, PE effort consists of three aspects: temporal, technical and cognitive effort. The temporal aspect, or PE time, can be seen to comprise the two other types of effort: the technical effort of performing the corrections, and the cognitive effort of identifying errors and planning corrections (Krings 2001, p. 178). These aspects of effort do not necessarily correlate: some errors in the MT may be easy to identify but require much editing, and conversely, other errors may be corrected with few keystrokes but involve considerable cognitive effort (Krings 2001; see also Koponen 2012). Approaches to measuring PE effort can be divided into product-based methods analyzing the final text produced by post-editing, and process-based methods investigating the process through which it was produced.

Product-based approaches to assessing PE effort generally involve comparing the MT output and PE version to determine the number and type of changes made by the post-editor using automatic edit distance metrics like HTER (Snover et al. 2006, 2009), which is then often used as an indicator of MT quality and PE effort. While HTER reflects PE effort to some extent, discrepancies between HTER scores and both perceived PE effort (Koponen 2012) and PE time (Koponen et al. 2012) have been observed. As Daems (2016, p. 118) points out, comparison of the MT and the final product of PE does not, for example, account for the fact that an editor may have returned to edit the same word or passage multiple times. While such process of editing a passage multiple times obviously involves increased technical effort, it may also reflect cognitive effort: according to Krings (2001, p. 530), needing to consider multiple potential translations of the same passage leads to increased cognitive effort. In general, cognitive effort is difficult to detect in the product. One approach is suggested by Choice Network Analysis (CNA, see Campbell 2000), where versions of the same text produced by different translators or post-editors are compared. Campbell (2000) argues that passages where the different versions are identical are cognitively less demanding than passages where translators produce multiple different translations, indicating that no one solution is obvious.

Process research methodologies offer a more detailed way to investigate PE processes and effort. Translation Process Research is an established branch of Translation Studies, with process studies being conducted since the 1980s (e.g. Lörscher 1986), first by using think-aloud methods (see Tirkkonen-Condit and Jääskeläinen 2000) and, since mid-1990s, using keylogging software such as Translog and, more recently, eye-tracking equipment (for an overview, see e.g. Göpferich et al. 2009; Hvelplund 2014; Carl et al. 2016). Keylogging provides more accurate information

particularly on the technical aspect of PE effort in terms of keystrokes and other editing operations (Daems 2016), but cognitive PE effort is again more difficult to measure directly. Pauses in keylogging data have been used as indicators of cognitive effort either in the form of extended pauses (e.g. O'Brien 2005, 2006; Koponen et al. 2012; Daems et al. 2015) or clusters of short pauses (Lacruz and Shreve 2014; Lacruz et al. 2014; Schaeffer et al. 2016) connected to a specific passage being edited. The definition of a "pause" varies: a survey of translation process research (Kumpulainen 2015) found that periods of inactivity lasting 1 s or 5 s were commonly used definitions, although both shorter and longer times occurred. Research on general writing processes also suggests that average pause length varies for different writers, and "short" or "long" pause should therefore be determined individually (Mutta 2016). Interpreting pauses and connecting them to specific MT errors or edits is also not straightforward (O'Brien 2006). As Englund Dimitrova (2005, 97) points out, cognitive processes during the pause may involve planning corrections to be carried out, or evaluating corrections that have already been carried out, or potentially something different. Methods to obtain information on cognitive processes during pauses include think-aloud protocols (e.g. Krings 2001, Vieira 2017b), and eye-tracking (e.g. Vieira 2014, 2017a; Daems et al. 2015), which relies on the assumption that the locations of gaze fixations indicate the focus of attention and fixation counts and duration reflect the amount of cognitive effort (see Rayner et al. 2012).

Combining error analyses and process-based effort methods, prior research mainly on SMT has linked increased PE effort with MT errors related to idioms and mistranslated words, reordering errors, omissions and extra words, syntax errors, and structural or coherence errors (Koponen 2012; Koponen et al. 2012; Lacruz and Shreve 2014; Lacruz et al. 2014; Popović et al. 2014; Daems et al. 2015), as well as specific source text features such as gerunds, consecutive noun phrases or prepositional phrases and word repetitions (O'Brien 2005; Aziz et al. 2014; Vieira 2014). So far, few studies have used process methods to examine PE effort involving NMT. Castilho et al. (2017, pp. 116–117) examined technical effort (keystrokes per segment) and temporal effort (seconds per segment) during PE and found that, in comparison to SMT, editing NMT involved less technical effort in all the language pairs studied, and less temporal effort in all but one language pair, although the differences are reported as marginal. Comparing process data in human translation to post-editing of both NMT and SMT output using a literary text, Toral et al. (2018) found that both MT approaches, and NMT in particular, reduced the number of keystrokes, and resulted in fewer but longer pauses compared to human translation. To the knowledge of the authors, no previous studies have investigated process data involving post-editing Finnish NMT.

2.3 PE output and quality

The use of PE output in evaluating MT generally relies on the assumption that edits reflect MT errors. However, although edits made during PE consist of changes the post-editor considered necessary, they do not necessarily always

involve actual errors. Studies have observed that post-editors sometimes over-edit, making "preferential" changes (de Almeida 2013, p. 100) in situations where the MT was already correct in terms of meaning and grammar. Preferential edits were found to account for between 16% and 25% of changes in de Almeida's study (2013). A pilot study involving English-to-Finnish PE data from five participants, reported in Koponen and Salmi (2017), assessed each edit in terms of correctness of meaning and grammar and necessity to make the MT sentence accurate and grammatical, and found that 38% of edits were unnecessary. Post-editors may also leave errors uncorrected or even introduce new errors. Depending on language pair, de Almeida (2013) observed that essential changes had not been made in 11% to 15% of cases, and errors had been introduced by the editor in 5% of cases. The analysis in Koponen and Salmi (2017) indicates that 3% of unedited words involved a case where a necessary edit had not been made, and 9% of edits were incorrect.

Potential errors or unnecessary edits are of course not exclusive to PE: similar tendencies have been observed in the context of revision of human translation. Arthern (1987) and Mossop (2018 and forthcoming) suggest rating scales for revisions containing categories for necessary and unnecessary interventions as well as errors or problems in the revised text left unnoticed or introduced by the reviser. Robin (2018) proposes a categorization considering the motivation for revisional modifications: Rule-based modifications, which relate to equivalence (of the source and target texts), linguistic rules or the translation brief, are deemed compulsory, while norm-based changes, motivated by norms of translation or the target language (on translation norms, see Toury 2012), and strategy-based changes, motivated by communication principles, are considered optional, and finally preference-based changes, which arise from the preferences of the individual reviser, are deemed "pointless", and may have no effect or even a detrimental effect on the text (Robin 2018, pp. 158–159). A parallel can be drawn to PE guidelines like TAUS (2010) and the International Standard ISO 18587, which generally advise that only minimal essential changes are made during PE, particularly in so-called light PE. We argue that studies using PE to evaluate and compare MT should take this tendency to over- or under-edit into account. The potential effect of preferential changes or editor errors can be mitigated by having multiple people carry out the PE task, as well as by an evaluation of the PE corrections themselves.

3 Experimental set-up: materials and methods

A pilot analysis of a subset of the data used here has been previously published in Koponen and Salmi (2017). In this article, we extend the analysis of PE changes and assessment of the correctness and necessity of the edits to the full dataset, using the same method as the previous article (Koponen and Salmi 2017, pp. 141–142). In addition, this article presents an analysis of the process data collected during editing.

3.1 Post-editing experiment

The PE data was collected from a total of 33 translation students on Master's level during two experiments in October 2016 (data set 1, 16 participants, P01–P16) and March 2017 (data set 2, 17 participants, T01–T17) during a PE task organized as part of two courses on MT and PE taught by the authors. The participants identified Finnish as their mother tongue (with the exception of two bilinguals, Finnish-Russian and Finnish-English) and their average age was 26.9 years (median 24). Although translation students may differ from more experienced professionals, the participants had experience of translation through their studies in a translator training program, and specific training in PE through the course during which the PE task was carried out, meaning that they can be considered semiprofessionals (see Englund Dimitrova 2005, p. 16). In the Finnish context, PE has not so far been commonly used (see Sect. 2.1), meaning few professional translators have training or experience in PE.

In the task, the students edited a short text machine translated from English into Finnish using three types of MT systems: RBMT, SMT and NMT. The English source text (ST) and the Finnish SMT and NMT versions were obtained from the ACL First Conference on Machine Translation (WMT16) News Task dataset (Bojar et al. 2016). The ST used in the PE experiment comes originally from the BBC website, and provides instructions for how to send material to the BBC. The text contains 27 sentences and a total of 385 words (for analysis of process data and specific features, the text was later further divided into 165 sub-segments, see Sect. 3.4). The goal of our PE experiment was to collect edits involving the same text from multiple participants to enable the comparison of how different editors correct the text, and to collect process data from their edits. For the purposes of the process data collection, we followed a common principle of translation process studies that the text needed to be short enough for the participants to complete in one session, for which reason only one text of short length was used. Furthermore, the text needed to be general enough that no specialized background knowledge was required from participants. The ST in question was selected because it was of general nature and contained no specialized terminology, and contained some repetitions that we considered interesting to be analyzed in terms of the editing process. The use of one relatively short text for analysis also enables a more detailed manual analysis of the edits, however, it obviously limits the generalizability of the results.

For the MT versions, outputs from two systems, AbuMaTran-NMT (Sánchez-Cartagena and Toral 2016) and UH Opus (Tiedemann et al. 2016), were selected from the WMT16 dataset from among the highest-ranking non-commercial systems. Since no RBMT system was included in the dataset, an RBMT version was produced using the online system Sunda.[1] Table 1 shows the number of words in each MT output and the TER score calculated against the WMT16 reference translation.

As PE choices made by individual editors may differ, it was considered important to collect editing data from each participant on each of the three MT outputs.

[1] http://www.sunda.fi/en/.

Table 1 Number of words in MT outputs and TER score against WMT16 reference translation

System	# Words in MT	TER
NMT (AbuMaTran-NMT)	277	76.95
RBMT (Sunda)	308	71.10
SMT (UH Opus)	252	92.86

Table 2 Division of the MT versions into sentence blocks for post-editing

Text block	Text version 1	Text version 2	Text version 3
Sentences 1–9	RBMT	SMT	NMT
Sentences 10–18	NMT	RBMT	SMT
Sentences 19–27	SMT	NMT	RBMT
Number of participants editing the MT version	13	9	11

Therefore, rather than having each participant edit a continuous text of one system output, the source text was divided into three blocks of nine sentences, and three text versions for editing were created combining blocks of MT output from each system. The system output used for each sentence block was rotated in the three text versions as shown in Table 2. Each participant was presented with one of these three versions for post-editing as a full text. The division of the MT outputs into three text blocks in each text version, the system output used for each block in the three text versions, and the number of participants editing each version is shown in Table 2. The rationale for this setting was to collect data involving all MT outputs from all participants and to enable comparison of different PE versions of the same sentences and to counteract a potential facilitation effect in the editing process. According to the facilitation effect, processing may become faster and pauses less frequent towards the end of the task because certain global decisions have been made at the beginning, recurring expressions are quicker to retrieve, and the growing text representation facilitates comprehension and production (see Englund Dimitrova 2005, p. 30).

For the PE task, the participants were given the following instructions (in Finnish), based on the principles of "light" PE as defined in the draft International Standard ISO/DIS 18587:2016[2]:

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

[2] The final standard has later been published in 2017 with some modifications, however, instructions used in the experiment correspond to the draft available at that time.

The PE task was carried out using Translog-II (Carl 2012), which collects keylogging data during the task. The participants were shown the entire text at the same time (ST on the left and MT on the right side of the screen) and were able to edit the sentences in any order they chose. Similarly to Vieira's (2017a, pp. 167–168) study, only the information available in the ST and MT were to be used in the task, no external sources of information like dictionaries were available. Although in a real-life PE situation external resources would be used, the focus of our experiment was to observe how the participants edit based on the information in the ST and MT alone, rather than their information search processes. Recording participant activity outside the Translog window would also have added a further complicating factor to the experimental setting. No time limit was set for the task, but the participants were advised to avoid spending excessive time on any one correction. In experiment 1, the SMI RED-m eye tracker was used to collect gaze data reflecting the overall reading process; however, as accuracy of the gaze data collected is not sufficient for analysis of fixations on specific words, it is excluded from the analysis presented in this paper.

3.2 Identifying PE changes

Following the PE product analysis started in Koponen and Salmi (2017), we used TER-plus (Snover et al. 2009) with the basic (H)TER parameters with stemming, synonymy and paraphrasing turned off to identify the word-level changes made by the participants in their edits. To further determine whether the substitutions involved actual word changes or changes to morphological form of the word, the MT and PE texts were also lemmatized using the morphological analyzer OMorFi (Pirinen 2008) and FinnPos morphological tagger toolkit (Silfverberg et al. 2015). Lemmatization was checked manually to detect potential errors in unknown words or homonymic words (for example, Finnish *tuo* can be either the nominal singular form of the pronoun *tuo* 'that', or second person singular imperative or third person singular indicative form of the verb *tuoda* 'bring').

MT and PE versions were aligned semi-automatically using the corrected lemmatization and the alignment data produced by TER-plus, and types of edits were categorized using the TER-plus edit operation annotations (insertion, deletion, substitution, shift or match). The alignment and annotations were again checked manually. In the manual phase, words with the same lemma were matched, and if the surface form differed, classified as changed word forms. Alignment and classification of substitutions, insertions and deletions were further checked. Rather than use the automatic alignment, where any deletion and insertion in the same position in the sentence are treated as a substitution, we considered only words with a semantic (or in some cases functional) equivalence to be substitutions. Correspondingly, cases where the participant added a word that had no equivalent in the MT version were classified as insertions, and cases where the participant removed a word from the MT leaving no equivalent in the PE version were classified as deletions. Furthermore, some cases of incorrect "matches", where homonyms appearing in the same sentence had been incorrectly aligned, were identified and changed to other

categories as applicable. Following this procedure, changes were annotated with one of the following categories according to the PE actions:

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

In some cases, the same word had been affected by more than one type of edit. Specifically, these involved cases where the word had both its form and position changed, or where the word was both substituted with another word and its position was changed. These cases were annotated with both categories ("form + order" or "word + order").

Our approach is similar to Bentivogli et al. (2016, 2018) in that we use classification of manual changes performed during PE to compare the NMT, SMT and RBMT systems. However, unlike Bentivogli et al. (2018) we conceptualize the classification in terms of *edits* rather than MT *errors* for the reason that not all changes made during PE necessarily involve actual MT errors (see Sect. 2.3). To determine which edits reflect actual errors, the correctness and necessity of each edit was further assessed as described in the next section.

3.3 Analysis of post-edits: correctness and necessity

After identifying the edit type, each word-level edit was assessed for correctness of meaning and language as well as for necessity of the edits, as in Koponen and Salmi (2017, p. 142). Correctness was in this connection assessed in terms of accuracy of meaning as well as the grammaticality of the target language (TL). Necessity was assessed based on whether the edits were essential to correct the meaning or language or whether they appeared to be preferential edits related to style or word choices. If the same word had undergone more than one type of edit, for example, a change of word form and a change of word order (see Sect. 3.2), both changes were analyzed separately for correctness and necessity.

As in Koponen and Salmi (2017), one PE text was chosen from each group of the three text versions edited, and it was assessed independently by each of the three authors of this article. The assessments were compared and differing decisions were discussed in order to agree on categorization (see example in Koponen and Salmi 2017, p. 142). At this point, a list of all the solutions accepted as correct for each word was created to serve as guideline for the rest of the assessment, and some general principles were determined. One such case involved translating the English second person forms, which in Finnish can be translated using either singular, plural, or the politeness form consisting of plural pronoun and singular verb form. A general decision was made that as the participants had been instructed to conduct light PE, all forms were accepted, and any changes of singular to plural or vice versa were

deemed a matter of style and therefore unnecessary. A similar decision was made on which edits could be considered as correct renderings for words that had been mistranslated in the MT output (such as "contribution" discussed in Sect. 4.3).

After agreeing on the assessment guidelines for all the three text versions, all edits in the remaining 30 texts were assessed by two different authors. Their inter-rater agreement on the assessment of each edited word was calculated for both correctness and necessity of the edits, per participant. On average, the agreement on the correctness of the edits was 95.34%, and the agreement on the necessity was 90.45%. The individual inter-rater agreement percentages per participant varied from 86.69 to 98% for correctness and from 86.4 to 93.95% for necessity. The averages indicate that there was more agreement on the correctness than on necessity. For determining the final results on correctness and necessity of the edits, one of the authors went through all the edits and harmonized the assessments in places where the two assessors had disagreed.

3.4 Analysis of process data: keystrokes and pauses as effort indicators

The process logs produced by Translog-II for each participant's session were first used to identify text production units. Following Carl et al. (2016, p. 35), a text production unit was defined as a sequence of continuous typing separated by pauses of 1 s or longer. This definition of pause length was selected as it is commonly used also in other studies (see Sect. 2.2) to separate continuous units of activity. Text production units identified according to this definition contain both text production or deletion and keyboard or mouse actions used to reposition the cursor. As the goal of this study was to further connect actual editing activity (deleting or inserting characters) to specific passages of the text, these units were then further divided by instances of cursor repositioning (mouse clicks, arrow keys or other function keys), to create units of text production or deletion activity involving a continuous passage in the text. These "editing units" used in our study therefore differ from the definitions of text production unit and activity unit as used by Carl et al. (2016) in that we use cursor repositioning in addition to pauses to divide the units.

For a more fine-grained identification of the passage affected by the production and deletion activity in each editing unit on a sub-sentence level, the 27 sentences in the ST and the MT versions were further divided into a total of 165 sub-segments containing meaningful units (mostly NPs and VPs and their subconstituents) according to their value from the perspective of the analysis. As the MT system outputs differed from each other, the segmentation was made in a way that enabled the annotation of different kind of expressions, as long as they represented a semantic or structural equivalent from the perspective of the translation. This segmentation was further used to identify specific sub-sentence level features in each MT output that were edited by all or nearly all participants who saw that version, as well as sub-segments left unedited by all or nearly all participants. Sub-segments edited by all but one participant were included as one participant might have overlooked an error. On the other hand, sub-segments edited by only one participant are likely to reflect individual preferences. The analysis of the PE changes described in 3.2 was further used

to describe typical PE changes for each MT in these sub-segments. Different PE outputs for some of the elements within these sub-segments were analyzed, although a full Choice Network Analysis (see Campbell 2000; O'Brien 2005) was not within the scope of this study.

After the sub-segment division, the Translog-II logs of participant sessions were annotated manually to indicate which sentence and which sub-segment within the sentence the editing activity affected. If multiple sub-segments were affected, the editing unit was annotated accordingly to indicate all sub-segments. Cases where typing during the editing unit did not lead to any visible changes in the final PE version were labeled separately. These cases involved, for example, adding a word which was later deleted, or deleting a word and then re-typing the same word.

To identify potential differences in PE effort between different MT outputs, effort indicators in the process data were compared. Technical effort was measured by calculating text producing or deleting keystrokes per ST word for each sentence and each participant. The number of ST words was used to enable comparison between systems, because the number of MT words may vary in different system outputs. A second indicator examined was the number of times a specific sub-segment within the sentence was visited by the same participant. The number of revisits is related to technical effort in that revising the same sub-segment multiple times can be assumed to lead to increased technical effort, but revising the same passage may also reflect cognitive effort (Krings 2001; see Sect. 2.2). To examine cognitive effort, we analyzed pauses between consecutive editing units affecting the same sentence. The rationale for focusing on only mid-sentence pauses was made because connecting pauses, in general, to either the activity happening after the pause (planning next edits) or before the pause (evaluating previous edits) is difficult (Englund Dimitrova 2005; see Sect. 2.2). Particularly between sentences from different MT outputs it is not possible to connect the pause to a specific system. Therefore, for the purposes of this analysis, we ignored pauses between different sentences (as well as pauses at the beginning and end of the session recording), while pauses occurring during the editing of one and the same sentence were assumed to be the most likely connected to that specific sentence. Using these mid-sentence pauses, we calculated the number of pauses per word and the average length of pauses per word, which previous studies have shown to be connected to cognitive effort (Lacruz and Shreve 2014; Lacruz et al. 2014).

The list of effort indicators used is the following

- Keystrokes (per word): total number of keystrokes used by each participant to edit the sentence in question, normalized by the number of ST words.
- Number of visits per sub-segment: the number of times each participant edited each sub-segment in the sentence.
- Number of mid-sentence pauses (per word): total number of pauses per participant between consecutive editing units affecting the same sentence, normalized by the number of words.
- Mid-sentence pause length (seconds per word): total length of pauses per participant between consecutive editing units affecting the same sentence, normalized by the number of words.

Table 3 Total number of words, number of unedited words and total number of edits

System output	NMT	RBMT	SMT
Total number of words in PE versions	3917	4389	3550
Total number of unedited words in PE	2250 (57.4%)	2417 (55.1%)	2051 (57.8%)
Total number of edits in PE	1667 (42.6%)	1972 (44.9%)	1499 (42.2%)

Table 4 Distribution of edit types (% of all edits) by system type

	NMT	RBMT	SMT
Inserted	357 (21.4%)	226 (11.5%)	**451** (30.1%)
Deleted	290 (17.4%)	**604** (30.6%)	132 (8.8%)
Form changed	493 (29.6%)	**618** (31.3%)	532 (35.5%)
Word substituted	**404** (24.2%)	374 (19.0%)	261 (17.4%)
Order changed	123 (7.4%)	**150** (7.6%)	123 (8.2%)
Total	**1667**	**1972**	**1499**

4 Results

4.1 Edit types and analysis of correctness and necessity of edits

Based on the analysis of PE changes identified in the final version produced by each participant (see Sect. 3.2), the distribution of different edit types was compared in NMT, RBMT and SMT versions. The total number of edits for each system was calculated based on the sum of all words edited by all of the participants who edited the MT version in question. Since the length of the versions produced by the different MT systems varied (see Table 1 in Sect. 3.1), the total numbers of words for each system also varies. Table 3 shows the total number of words analyzed as the total of all PE versions, the number of unedited words and the number of edits. As Table 3 shows, PE versions of the RBMT output had the largest number of total words, as well as the largest number and percentage of edits.

The distribution of edit types as well as the number of unedited words per system is shown in Table 4. As noted in Sect. 3.2, in some cases a word order change was combined with a word form change or word substitution. These cases were treated as two separate edits.

From Table 4 we can see that word form changes are, in general, the most common type of change. The highest number of word form changes occur in the RBMT output. Word substitutions are also common for all systems, with the highest number occurring in NMT output. Word order changes are relatively uncommon in the sentences analyzed, with the highest number occurring in the RBMT output. Comparing the systems, differences can be observed particularly in the numbers of inserted and deleted words. The RBMT version has more than double the number of deletions than the other two versions combined, while deleted words are least common in SMT. Conversely, the SMT output involves

Table 5 Correctness and necessity analysis, divided by MT system and edit type

Edit type	Total	Correct-necessary	Correct-unnecessary	Incorrect-necessary	Incorrect-unnecessary
NMT					
Inserted	357	244 (68%)	93 (26%)	8 (2%)	12 (3%)
Deleted	290	147 (51%)	110 (38%)	13 (4%)	20 (7%)
Form changed	493	313 (63%)	158 (32%)	14 (3%)	8 (2%)
Word substituted	404	243 (60%)	142 (35%)	13 (3%)	6 (1%)
Order changed	123	45 (37%)	73 (59%)	4 (3%)	1 (1%)
Total edits	1667	992 (60%)	576 (35%)	52 (3%)	47 (3%)
RBMT					
Inserted	226	129 (57%)	83 (37%)	5 (2%)	9 (4%)
Deleted	604	121 (20%)	456 (75%)	4 (1%)	23 (4%)
Form changed	618	205 (33%)	386 (62%)	10 (2%)	17 (3%)
Word substituted	374	152 (41%)	190 (51%)	19 (5%)	13 (3%)
Order changed	150	60 (40%)	85 (57%)	3 (2%)	2 (1%)
Total edited	1972	667 (34%)	1200 (61%)	41 (2%)	64 (3%)
SMT					
Inserted	451	333 (74%)	97 (22%)	12 (3%)	9 (2%)
Deleted	132	56 (42%)	53 (40%)	9 (7%)	14 (11%)
Form changed	532	392 (74%)	118 (22%)	17 (3%)	5 (1%)
Word substituted	261	161 (62%)	71 (27%)	20 (8%)	9 (3%)
Order changed	123	67 (54%)	54 (44%)	2 (2%)	0 (0%)
Total edited	1499	1009 (67%)	393 (26%)	60 (4%)	37 (2%)

the highest number of words inserted by the participants. A Chi squared (χ^2) test shows that the differences in distribution are statistically significant ($p < 0.001$). The contributing factors mainly appear to be a strong positive association between RBMT and deleting words, and negative association between RBMT and inserting words, as well as a positive association between SMT and inserting words and strong negative association between SMT and deleting words.

To examine the extent to which these edits represent MT errors, the correctness and necessity of each edit was also assessed. Table 5 shows the results of the correctness and necessity analysis for different edit types, compared between NMT, SMT and RBMT. The columns show the total numbers of each edit type and the classification of these edits into four categories: correct and necessary edits (MT error successfully corrected), correct but unnecessary edits (no error in the MT version), incorrect but necessary edits (error in the MT but PE correction is also incorrect), and finally incorrect and unnecessary edits (no error in the MT, but the PE version introduces an error). The total number of edits in each category and the percentage of all edits is shown for each category.

Although the number of edits is highest for RBMT, the assessment of necessity indicates that most edits to the RBMT output are in fact unnecessary even

if correct (61%). In the case of the NMT and SMT outputs, on the other hand, most edits are both correct and necessary (60% and 67%, respectively). This suggests that the number of errors corrected (as opposed to unnecessary edits) in the sentences analyzed is in fact lowest for RBMT and highest for SMT. This is most evident in the case of deleted words (75% of the deletions in RBMT sentences are correct but unnecessary), and word form changes (62% correct but unnecessary). Deletions also appear to lead to a relatively high proportion of incorrect edits, particularly in the SMT sentences, where 11% of deletions are both incorrect and unnecessary, and 7% represent cases where the MT is incorrect but deletion did not correct the problem. In both NMT and RBMT, more than half of word order changes are also unnecessary, whereas for SMT, more than half of word order changes are necessary. According to a χ^2 test the difference in distribution is statistically significant ($p < 0.001$). The major contributing factor appears to be the number of correct but unnecessary edits in the RBMT sentences. Unnecessary edits in the RBMT output appear to be connected to two specific features, 2nd person forms and subject pronouns (see Sect. 4.3). Comparing the numbers of necessary changes, the NMT output contained the smallest number of word order errors but the highest number of lexical errors (substitutions). Compared to SMT, the NMT output also contained fewer omissions (based on words inserted by participants) and fewer word form errors, but a larger number of extra words (based on words deleted by participants). On the other hand, the RBMT output contained even fewer word form errors and omissions than NMT.

Overall, the overwhelming majority of changes made by the participants to all MT versions are correct; only 5% to 6% of the changes are assessed as incorrect. In addition to edited words, the correctness of unedited words was assessed to determine whether the PE versions still contained errors that had been left uncorrected by the participants. The number of identified cases where some type of edit would have been necessary was 49 in NMT sentences (2.2% of all unedited words), 56 in SMT (2.7%) and 80 in RBMT sentences (3.3%). These instances of overlooked necessary corrections included errors present in the MT which had been left uncorrected, but also some cases where a change made by the participant elsewhere in the sentence would have necessitated also changing another word (or several words) but these had been left unchanged. An example is changing the translation of "you" from plural to singular in one part of the sentence but neglecting to change another part accordingly. The small number of incorrect edits and missed errors observed points to good overall quality of the edits, despite the participants being translation students.

4.2 PE effort metrics in the process data

Overall, variation was observed between the process data of the participants. Total PE time varied from 5 min 37 s to 39 min, the number of text producing or deleting keystrokes varied from 113 to 3347, and average pause time varied from 2.6 to 7.4 s. To examine potential differences in the amount of PE effort when editing sentences from the three different systems, we compared the following PE effort

indicators: the number of keystrokes per word, the number of visits to the same sub-segment (by the same participant), the number of pauses per word between consecutive units affecting the same sentence, and the average length of pauses (seconds per word) while editing a continuous passage. The number of keystrokes reflects technical effort, while the number and length of pauses reflect cognitive effort, and the number of visits may reflect both. These indicators were compared in each of the three text blocks consisting of nine sentences from the same MT output (see Table 2 in Sect. 3.1). Table 6 shows the average number of keystrokes per word, average number of visits per sub-segment, average number of pauses per word and average length of pauses in seconds per word for each system and each text block. The highest mean value for each indicator and text block is bolded in Table 6.

Differences observed between the systems appear to be focused on specific text blocks. In text blocks 1 and 2, the number of keystrokes is highest for NMT and lowest for RBMT. The difference between these two outputs is statistically significant (text block 1 $p < 0.005$, text block 2 $p < 0.01$), while differences between the NMT and SMT outputs or RBMT and SMT are not. The number of times each sub-segment was visited appears to be overall lowest for the NMT output. The only statistically significant difference is observed in text block 3, where RBMT has a higher average number of visits than the other two systems ($p < 0.005$). The average number and length of pauses is highest for the NMT output in text block 1, for SMT in text block 2, and for RBMT in text block 3. For both number and length of pauses, the difference is statistically significant only in text block 3 ($p < 0.001$).

Although the differences are small, they suggest that in text block 1, edits to the NMT output involved slightly more keystrokes (technical effort) and pauses (cognitive effort) than the other two systems, while in text block 2, NMT output involved more keystrokes but fewer and shorter mid-sentence pauses than SMT, suggesting the edits in SMT involved more cognitive effort. In text block 3, both RBMT and SMT outputs involved more keystrokes and visits to the same sub-segment, and particularly the number and length of mid-sentence pauses indicate potentially more cognitive effort for RBMT. Some of the differences when comparing values in the three text blocks may be due to the facilitation effect: toward the end of the task (text block 3) the participants were able to proceed faster because they were already more familiar with the text topic and had made some global decisions regarding recurring features.

4.3 Examples of typical edit types for NMT, RBMT and SMT

Although a detailed analysis of the specific errors or features in each system output is not within the scope of this article, some examples of potential problems are examined here based on the identification of sub-segments edited by all participants or all but one participant. Sub-segments edited by all but one are included to account for situations where one participant may have missed an error. Cases not requiring editing also identified as sub-segments not edited by any participant, or only one participant, to account for edits due to individual preference. Table 7 shows the numbers of sub-segments (out of a total of 165, see Sect. 3.4) edited by all or

Table 6 Effort indicators in process data by text block and system

Text block	System output	Keystrokes (per word)			Visits (per sub-segment)			Pauses (per word)			Average pause length (seconds per word)		
		Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
1 (sent. 1–9)	NMT	**3.03**	3.00	2.38	1.82	1.00	1.37	**0.48**	0.38	0.38	**1.82**	1.27	1.70
	RBMT	2.13	1.44	1.91	**1.96**	1.00	1.58	0.38	0.33	0.27	1.25	1.00	1.01
	SMT	2.40	1.69	1.97	1.85	1.00	1.18	0.37	0.33	0.25	1.51	1.08	1.49
2 (sent. 10–18)	NMT	**3.17**	2.75	2.60	2.09	1.50	1.65	0.51	0.36	0.42	1.59	1.07	1.66
	RBMT	2.48	1.95	2.14	**2.11**	2.00	1.48	0.46	0.40	0.36	1.83	1.12	2.18
	SMT	2.87	1.92	3.13	2.04	2.00	1.50	**0.54**	0.38	0.43	**2.00**	1.31	2.12
3 (sent. 19–27)	NMT	2.79	1.92	2.80	1.90	2.00	1.18	0.40	0.29	0.34	1.30	0.97	1.11
	RBMT	**3.06**	2.64	2.11	2.09	2.00	1.44	**0.55**	0.46	0.32	**1.89**	1.48	1.49
	SMT	**3.06**	2.40	2.49	**2.40**	2.00	1.64	0.49	0.34	0.56	1.40	0.99	1.48

Table 7 Number of sub-segments edited by all, nearly all, none, or nearly none of the participants

	NMT	RBMT	SMT
Sub-segments edited by all participants	27	29	20
Sub-segments edited by all but one participant	15	24	32
Sub-segments edited by none of the participants	48	33	46
Sub-segments edited by one participant	30	35	35

nearly all participants who saw the text version in question, and sub-segments edited by none or nearly none of the participants.

Compared across systems, SMT has the smallest number of sub-segments edited by all (20 out of 165), and RBMT the highest number (29). When sub-segments edited by all but one participant are included, all three systems are on the same level (NMT 42 in total, RBMT 44 and SMT 41). NMT also has the largest number of cases where no participant edited the sub-segment or only one did (71 in total), followed by SMT (63), while RBMT (52) has the smallest number of unedited sub-segments. Based on these comparisons, NMT appears most successful, particularly in that it contains the largest number of sub-segments not requiring editing. We identified 14 cases where none of the participants had edited a specific sub-segment in any of the MT versions. Nearly all (11 out of 14) cases consisted of punctuation marks at the end of a sentence, and these sub-segments were excluded from further comparisons. One case was also observed where a sub-segment was edited by only one participant out of the total of 33, indicating a clearly individual preference. This involved changing the translation of the word "how", which all three system outputs translated as *miten*, to the synonymous *kuinka.*

Some general observations can be made about typical edits for specific systems. The high number of unnecessary word form changes and deletions as well as unnecessary changes in the RBMT output (see Table 5 in Sect. 4.1) can be mostly explained by changes to personal and possessive pronouns. Firstly, RBMT had rendered the 2nd person pronouns (*you, your*) occurring in the ST consistently by using the plural form (*te, teidän*). The SMT output consistently contains the singular (*sinä, sinun*), while the NMT output varies. As all participants had sentences from each MT output, the singular and plural forms vary in each text version. Examples of edits made include changing these forms to be consistent throughout the text. While changing plural forms to singular appeared more common, even in the version where the RBMT output was presented first, changes in both directions (plural to singular and singular to plural) were observed. A detailed analysis of the changes is, however, not within the scope of this article. In addition, participants deleted personal pronoun subjects or possessive pronouns preceding a noun: in Finnish, pronominal subjects can be omitted in 1st and 2nd person, since they are signaled by the verb form, and personal pronoun possessives are similarly redundant before a noun containing the possessive ending (for example *(minun) kuva + ni* '(my) picture + POSS-1SG' versus *(sinun) kuva + si* '(your-SG) picture + POSS-2SG'). The RBMT output included the

pronominal subject in all 31 cases where one occurred in the ST, and 17 out of 18 cases of possessive pronouns. In contrast, the NMT output contained only 4 pronominal subjects and 1 possessive, and SMT output contained only 5 pronominal subjects and 1 possessive pronouns. While both changes (plural vs singular form of you, omission of pronouns) are correct and may improve style, they are not required for either meaning or grammatical correctness, and therefore unnecessary for light PE.

Ambiguity resulting from sentence-initial imperative verb forms was also observed to cause problems for both NMT and SMT, while RBMT generally rendered imperative verbs forms correctly. Common errors involved translating an imperative verb form, for example in the sub-segment "Email us", with a noun, *Sähköpostilla* 'by email' (email-SG-ADE), which makes the sentence ungrammatical. In the same sub-segment, RBMT included the imperative verb form *Lähettäkää* 'send-2PL-IMP'. Other errors involving imperatives in NMT and SMT output included incorrect verbs and omissions. Omissions of words were, in general, most common in SMT output and to a lesser extent in NMT, while the RBMT rarely omitted words.

Other verb forms such as the ing-participle forming the continuous aspect of verbs also caused problems in both the NMT and SMT outputs. Common errors included translating these verb forms as infinitives or nouns where Finnish would use the simple present tense verb form. Such errors make the sentence ungrammatical, as in Example 1, which shows the source sentence, the different MT output versions and a gloss of the MT.

Example 1

ST	Is something significant, bizarre or unusual happening where you live?
NMT	Onko jotain merkittävää, outoa tai epätavallista tapahtua siellä, missä asuu?
Gloss	'*Has something significant, bizarre or unusual happen there, where lives?*'
SMT	Onko jotain merkittävää, outoja tai epätavallisia tapahtumia, missä asut?
Gloss	'*Are there something significant, bizarre or unusual happenings, where you-SG live?*'
RBMT	Tapahtuuko jotakin merkittävää, eriskummallista tai epätavallista, missä te asutte?
Gloss	'*Does something significant, bizarre or unusual happen, where you-PL live?*'

Both the NMT and SMT systems start the sentence with *onko* (*on* 'be-3SG' + interrogative particle *ko*) but mistranslate the participle "happening" either as infinitive *tapahtua* 'to happen' (NMT) or as a plural partitive form of the noun *tapahtumia* 'events' (SMT). The RBMT system renders the structure correctly by moving the main verb to the beginning of the sentence and generating a present tense question form *tapahtuuko* (*tapahtuu* 'happen-pres-3SG' + *ko*).

Instances of untranslated words, which other studies have observed in translations of all MT types, were rare in the sentences analyzed in this study. The SMT output contained only one term, "breaking news", left untranslated. The NMT output contained one sentence with several untranslated words, as shown in Example 2. The translated words are underlined in Example 2, although **uutise* is not a correct form of the word *uutinen* 'news'.

Example 2

ST	The part you play in making the news is very important.
NMT	The osa of of the uutise is very tärkeä.

Two sub-segments appeared to cause difficulties in all MT outputs, as they have been edited by all participants. The first case was "your Twitter username". In both SMT and RBMT output, "username" was translated using a correct word but incorrect grammatical form. NMT rendered the sub-segment as *Twitteriin pukeutunutta* 'dressed in Twitter', making the segment unintelligible. The second case was the sub-segment "unless you ask us not to", which occurred three times in the text. The RBMT output consistently used the expression *jos te ette kiellä meitä* ('if you-PL do not forbid us'), which is correct in terms of both language and meaning. In one instance, the NMT output rendered the verb "ask" incorrectly as *et kysy* ('do not inquire'), but in other cases the NMT and SMT outputs contained varying translations of "ask" using forms of the verb *pyytää* 'request' but omitting the final "not to" leading to translations like "unless you request us", as shown in Example 3.

Example 3

ST	We will publish your name as you provide it (unless you ask us not to) […]
SMT	Julkaisemme nimesi kun se (ellet pyydä meitä) […]
Gloss	*[We] publish [your-SG] name when it (unless [you-SG] request us) […]*

Here, the SMT system also omits any translation for "you provide", leaving the MT version ungrammatical and omitting information. In general, omissions of this type were most common in SMT output and to a lesser extent observed also in NMT, while the RBMT rarely omitted words. (For a more detailed comparison of the edits of the NMT output of this sub-segment, see Koponen and Salmi 2017, pp. 145–146.)

Specific words with no obvious translation in Finnish also caused difficulties across in all three outputs. One example is "contribution", which occurred four times, referring to text or pictures sent to BBC. The MT outputs contained four semantically different options implying payment, verbal contribution to a discussion, or more generally, something contributed. Only the last option (using the word *panos*), which the RBMT output uses in three of the four cases and the SMT and NMT outputs each use once, can be considered acceptable, although it generally carries a connotation of considerable effort and is therefore stylistically not appropriate. The difficulty in translating this word led to considerable variation also in the PE outputs (altogether 17 different variants provided by all participants in the four sub-segments).

5 Discussion

Previous studies comparing NMT to SMT have suggested both an overall reduction of errors as well as a reduction in specifically morphological errors and word order errors in various language pairs (Bentivogli et al. 2018; Klubička et al. 2017, 2018; Popović 2018). In the language pair English-to-Finnish, Toral and

Sánchez-Cartagena (2017, p. 1070) reported that NMT reduced both inflection and reordering errors approximately 12% compared to SMT. In our study, the total number of word order edits was the same for the NMT and SMT outputs (see Table 4), however, comparing the number of necessary edits (Table 5), NMT can be seen to reduce word order errors compared to both the SMT and RBMT outputs. In the case of word form changes, the NMT output involved fewer necessary edits than the SMT but more than the RBMT output. Examination of specific example sub-segments suggest certain verb forms like the continuous aspect and imperative forms to be problematic for both NMT and SMT, while the RBMT generally contained correct forms in these instances (see Sect. 4.3). This is in line with Burchardt et al. (2017), who observe that the RBMT system was more successful in producing the correct verb tense, aspect and mood than either NMT or SMT, and with observations on ambiguity in the study by Popović (2018). In previous studies, lexical errors, mistranslations and omissions have been observed to be common in NMT output (e.g. Castilho et al. 2017; Klubička et al. 2018; Toral and Sánchez-Cartagena 2017). A similar observation can be made in our analysis, where the number of necessary word substitutions is highest in NMT output. The number of necessary insertions suggests that the NMT output in our study involved fewer omissions than the SMT output, although more than RBMT. Compared to the SMT and RBMT outputs, the NMT output also involved more necessary deletions, indicating extra words.

Some studies (e.g. Castilho et al. 2017) have suggested that reductions of word order and word form errors improve the fluency of NMT, but not necessarily the adequacy, and that this improving fluency may become misleading as NMT potentially generates grammatically correct sentences which do not correspond to the meaning of source text. Such tendency was not observed in our study, although some examples of grammatically correct and fluent passages with incorrect meaning were found. If the participants had been misled by the NMT output, the number of missed errors (meaning cases where a correction would have been necessary but no edit was made) would be expected to be higher. On the contrary, only 2.2% of unedited words in the NMT output were categorized as missed errors, compared to 2.7% in the SMT output and 3.3% in the RBMT output. However, the analysis in this study is based on relatively small-scale data from a short text, which naturally limits the generalizability of the results.

Overall, the analysis of correctness and necessity of the edits shows that unnecessary changes were common even though the participants had been instructed to focus on errors in meaning or grammar and avoid changes intended to improve the style or fluency of the text. This observation is in line with de Almeida's (2013) assessment of post-editor corrections. In particular, a much higher number of correct but unnecessary changes were made to the RBMT output (61%) compared to NMT (35%) and SMT (26%). As discussed in Sect. 4.3, most of this difference is explained by changes to the translation of 2nd person forms and the deletion of redundant personal pronouns as subjects or possessives in the RBMT output. In the context of light PE, in particular, such changes can be considered inefficient, and a higher number of unnecessary changes also increases the risk of introducing errors and leading to lower PE quality (see Vieira 2017a). While some of the unnecessary changes may be preferential (cf. de Almeida 2013; Robin 2018), Robin's (2018, p.

159) framework of revision modifications suggests another reason for cases where all or nearly all participants made unnecessary changes: they may be classified as strategy-based modifications made to improve the readability of the text. The difference in the relative numbers of unnecessary changes may therefore indicate differences in readability or fluency of the MT outputs. Fluency is a known issue for RBMT, and in the case of Finnish, frequent use of redundant pronouns, for example, may contribute to an overly literal and stilted "machine-translationese" style. Less frequent unnecessary changes in NMT (but also SMT) may therefore point to better fluency.

The results for effort indicators in the PE process data somewhat correspond to Castilho et al. (2017, pp. 116–117), who observed that technical and temporal effort was lower for NMT than SMT in all language pairs but one (English-to-Russian). In our sub-sentence level analysis of edited versus unedited sub-segments (Sect. 4.3), a larger number of sub-segments in the NMT output were left unedited by all or nearly all of the participants, which is in line with Castilho et al. (2017). The NMT output also contained a higher number of sub-sentence level segments which were edited in one pass without revisits (Sect. 4.2). As mentioned above, in order to involve all participants to edit all three MT outputs, each participant edited a text combining blocks of MT output from each MT system. With regard to technical effort, editing the NMT output involved more keystrokes per word than the RBMT and SMT outputs in text blocks 1 (sentences 1–9) and 2 (sentences 10–18), but fewer keystrokes in text block 3 (sentences 19–27), with statistically significant differences between NMT and RBMT in text blocks 1 ($p < 0.005$) and 2 ($p < 0.01$). The NMT output also contained a higher number of sub-sentence level segments edited in one pass without revisits, which was particularly evident in text block 3 ($p < 0.005$). These two observations suggest that more technical effort was involved in editing the NMT output in text blocks 1 and 2, but less in editing text block 3 compared to the other outputs. With regard to cognitive effort, Toral et al. (2018) reported longer but fewer pauses when editing NMT compared to SMT. In our study, the average number and length of pauses differed in the text blocks. Both measures were highest for the NMT output in text block 1, but for SMT in text block 2 and for RBMT in text block 3. Only the differences in text block 3 were statistically significant ($p < 0.001$). In text block 3, RBMT also had a higher average number of visits to the same sub-segment than the other two systems ($p < 0.005$). This indicates potentially increased cognitive effort when editing the RBMT output, and decreased cognitive (as well as technical) effort when editing the NMT output of text block 3.

Although differences were observed between the three outputs in the product-based analysis of edit types as well as correctness and necessity of edits, the process measures are less clear. This may be affected by the process data collection set-up, where the participants saw the entire text, containing sentences from all three MT outputs, at once. As noted by Vieira (2017b, p. 102), in this format, participants do not approach the text as isolated sentences; rather, they are able to plan an overall strategy (such as deciding on a consistent 2nd person form) and conduct edits in multiple rounds. This behavior was indeed observed for most of our participants: only two of them went through the text in one pass from beginning to end, while all others exhibited some form of backtracking and revising their own edits.

Some overall planning is likely to have happened during the initial pauses (mean: 40.39 s) before the first editing operation. Some planning of the corrections to a given sentence probably also happened during pauses when switching to a new sentence, which are on average (mean: 10.35 s) longer than mid-sentence pauses (mean: 3.74 s). This planning is not captured by the focus on mid-sentence pauses used in this study, but we argue that focusing on mid-sentence pauses allows us to connect the effort reflected by the pause to a specific passage in the text with more certainty, as pauses between sentences may also be connected to evaluating the previous sentence or text produced so far. A facilitation effect was also observed in that recurring expressions toward the end of the text were corrected faster, often by copying a previous solution. In the comparison of the MT outputs, this effect is mitigated by the fact that participants edited the MT outputs in different order.

Although a detailed comparison of the different participants is outside the scope of this paper, it should be noted that considerable variation was observed between the participants in terms of both the product and process-based metrics. As noted in Koponen and Salmi (2017, p. 144), one participant (identified as "E") assessed in the pilot stage had more instances of necessary changes not performed than the other four combined, and the full analysis reported in the present paper identified another participant with an even higher number of overlooked errors. The number of unnecessary changes also varied greatly between participants. Although using PE offers a way to provide information about MT errors, this reinforces that the assumption that edits represent errors is not unproblematic. As in previous studies (e.g. Koponen et al. 2012; Toral et al. 2018), effort indicators in process data varied greatly between the participants. The participants in this study were translation students, which may affect the PE quality and processes compared to more experienced professionals. However, our participants had training in both translation and specifically in PE, meaning that they can be considered semiprofessionals (see Englund Dimitrova 2005, p. 16), and no significant differences have been found between experienced translators and novices in a previous study (Guerberof Arenas 2014). As observed in Sect. 4.1, the small number of incorrect edits also points to overall good quality of the student edits. The use of student data can be further justified by the fact that PE is not yet a common practice in the Finnish translation field, and professional translators with PE experience would therefore be difficult to find. As MT and PE are increasingly integrated into the workflows, we are planning further studies with professional translators and more experienced post-editors.

6 Conclusion and future work

This paper presents a comparison of PE changes performed on NMT, RBMT and SMT output edited by a total of 33 translation students acting as participants in an English-to-Finnish PE experiment. Combining a product-based and a process-based approach, the objectives of our analysis were (1) to identify potential differences in the types of edits performed, and (2) to identify differences in the effort indicators in PE process data. Based on these differences, our aim was to provide information about the number and type of errors and their effect on PE effort in the output of

these three MT system types particularly in a morphologically rich target language like Finnish.

Our product-based analysis of PE changes shows that the NMT output contains a larger number of sub-sentence level sub-segments left unedited than either RBMT or SMT, although the total number of edited words is higher for NMT than SMT (but lower than RBMT). A further assessment of the correctness and necessity of the edits revealed that a considerable number of edits involved correct but unnecessary changes; in particular, RBMT involves a higher proportion of unnecessary changes. Statistically significant differences in the distribution of edit types between NMT, SMT and RBMT were also observed, the most prominent factor being the high number of deletions in the RBMT and the high number of insertions in the SMT outputs. Comparison of necessary changes, which can be considered to reflect MT errors rather than preferential changes, suggests that the NMT output contained the smallest number of word order errors but the highest number of lexical errors (substitutions). Compared to SMT, the NMT output also contained fewer omissions (based on words inserted by participants) and fewer word form errors, but a larger number of extra words (based on words deleted by participants). On the other hand, the RBMT output contained even fewer word form errors and omissions than NMT. As in Popović (2018) and Burchardt et al. (2017), problems related to certain verb forms and ambiguity were observed for NMT, while RBMT was more likely to handle them correctly.

It should be emphasized that the comparison of edits is based on only one short text passage of 27 sentences and only three systems outputs, which naturally limits the generalizability of the results. Some caution is therefore necessary in interpreting the results as errors typical to NMT, RBMT or SMT. A short text was selected due to the process data collection set-up of this study, where the use of a longer text or multiple texts to be post-edited by the participants was not deemed feasible due to practical time limitations. A manual analysis of a short text was beneficial in enabling a detailed comparison of multiple participants' choices in editing the passage, and the comparison of edit patterns related to each of the three MT outputs used in the study suggest differences in the types of edits the participants found necessary. However, larger scale studies involving larger corpora and more varied text types would be necessary to test whether, and to what extent, these observations can be generalized. A further limitation is acknowledged in that NMT has undergone rapid developments very recently, and although the NMT system used in this study represented state-of-the-art at the time of data collection, current systems might produce different results. We therefore plan to carry out further experiments with more advanced systems.

Results of the process-based comparison of effort indicators are less conclusive, although some differences were observed related to specific blocks of sentences in the text edited. Based on the average number of keystrokes per word, editing the NMT output appeared to involve less technical effort than the other two outputs in two of the text blocks, where differences were observed to be statistically significant, but less technical effort in the third. Particularly in the third text block, where more sub-sentence level segments in the NMT output were also edited in one pass. Based on the average number and length of pauses, editing the RBMT appeared to involve

increased cognitive effort in the third block of sentences, with a statistically significant difference compared to the other two outputs in this block. The process-based analysis is limited by the experimental setting where participants saw a full text version which contained output from all three systems. The benefits of this approach are that it allows for collecting data from the same participant editing all the MT systems, and that it is more conducive to text-level editing approaches (see Vieira 2017b, p. 102). As a trade-off, the set-up complicates connecting effort indicators to specific sentence or MT output. Future work is planned to modify the setting so that each participant edits output from different systems as separate tasks. For a closer analysis of cognitive effort, use of gaze data for fixation analysis is also planned in future work.

Despite these limitations, we believe the current paper contributes to the understanding of differences between NMT, RBMT and SMT approaches particularly in a morphologically rich language. Future work could include a detailed analysis on the linguistic features in the edits such as the changes described in Sect. 4.3, as well as a full Choice Network Analysis on the edits mentioned in Sect. 3.4. Further work involving larger datasets and more varied texts remains to be done to determine whether differences in edit types can be observed on a more general scale, and to determine whether and how differences in edit types are reflected in cognitive effort indicators. Furthermore, our findings regarding the variation in both the correctness and necessity of PE changes and in process data have broader implications for the use of PE data for MT error evaluation. We suggest that studies using PE as an evaluation method should also include some assessment of the PE quality to identify both potential participant errors and preferential edits and take into account individual differences in process metrics.

Acknowledgements Open access funding provided by University of Turku (UTU) including Turku University Central Hospital. The authors wish to thank all the students who participated in the PE process experiment. We also thank Dr. Miikka Silfverberg for assistance with use of the OMorFi and FinnPos tools and the two anonymous reviewers for their valuable comments.

References

Arthern PJ (1987) Four eyes are better than two. In: Picken C (ed) Translating and the Computer 8. Proceedings of a conference held on 13–14 November 1986. Aslib, London, pp 14–26

Aziz W, Koponen M, Specia L (2014) Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In: O'Brien S, Balling LW, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and application. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 170–199

Bentivogli L, Bisazza A, Cettolo M, Federico M (2016) Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP 2016), Austin, pp 257–267

Bentivogli L, Bisazza A, Cettolo M, Federico M (2018) Neural versus phrase-based MT quality: an in-depth analysis on English–German and English–French. Comput Speech Lang 49:52–70

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), Association for Computational Linguistics, Stroudsburg, pp 131–198

Burchardt A, Macketanz V, Degdari J, Heigold G, Peter J-T, Williams P (2017) A linguistic evaluation of rule-based, phrase-based, and neural MT engines. Prague Bull Math Linguist 108:159–170

Campbell S (2000) Critical structures in the evaluation of translations from Arabic into English as a Second Language. 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, European Language Resources Association (ELRA), pp 4108–4112

Carl M, Schaeffer M, Bangalore S (2016) The CRITT translation process research database. In: Carl M, Bangalore S, Schaeffer M (eds) New directions in empirical translation process research: exploring the CRITT TPRDB. Springer, Cham, pp 13–54

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017) Is neural machine translation the new state of the art? Prague Bull Math Linguist 108:109–120

Crego JM, Kim J, Klein G, Rebollo A, Yang K, Senellart J, Akhanov E, Brunelle P, Coquard A, Deng Y, Enoue S, Geiss C, Johanson J, Khalsa A, Khiari R, Ko B, Kobus C, Lorieux J, Martins L, Nguyen D, Priori A, Riccardi T, Segal N, Servan C, Tiquet C, Wang B, Yang J, Zhang D, Zhou J, Zoldan P (2016) Systran's pure neural machine translation systems. CoRR arXiv:1610.05540. Accessed 27 Oct 2018

Daems J (2016) A translation robot for each translator? A comparative study of manual translation and post-editing of machine translations: process, quality and translator attitude. Dissertation, Ghent University

Daems J, Vandepitte S, Hartsuker R, Macken L (2015) The impact of machine translation error types on post-editing effort indicators. In: Proceedings of MT summit XV: Fourth Workshop on Post-editing Technology and Practice (WPTP 4), pp 31–45

de Almeida G (2013) Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages. Dissertation, Dublin City University

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

Ervasti H (2017) MT in the Translation Industry. Paper presented at the Second Finnish Workshop on Machine Translation (University of Helsinki, 1 November 2017) http://blogs.helsinki.fi/language-technology/files/2017/09/FINMT2017-Ervasti.pdf. Accessed 27 Oct 2018

European Commission (2018) eTranslation. A Connecting Europe Facility website. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation. Accessed 27 Oct 2018

Göpferich S, Jakobsen AL, Mees IM (eds) (2009) Behind the mind: methods, models and results in translation process research, vol 37. Copenhagen Studies in Language Series. Samfundslitteratur, Copenhagen

Green S, Chuang J, Heer J, Manning CD (2014) Predictive translation memory: a mixed-initiative system for human language translation. Proceedings of the 27th annual ACM symposium on user Interface Software and Technology, pp 439–448

Grönroos SA, Virpioja S, Kurimo M (2017) Extending hybrid word-character neural machine translation with multi-task learning of morphological analysis. In: Proceedings of the conference on machine translation (WMT), volume 2: Shared Task Papers, pp 296–302

Guerberof Arenas A (2014) The role of professional experience in post-editing from a quality and productivity perspective. In: O'Brien S, Balling LW, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 51–76

Hurskainen A, Tiedemann J (2017) Rule-based machine translation from English to Finnish. In: Proceedings of the conference on machine translation (WMT), volume 2: Shared Task Papers, pp 323–329

Hvelplund KT (2014) Eye tracking and the translation process: reflections on the analysis and interpretation of eye tracking data. In: Martin RM (ed) MonTI Special Issue 1: Minding translation, pp 201–223

ISO/DIS 18587 (2016) Translation services-post-editing of machine translation output-requirements. International Organization for Standardization

Junczys-Dowmunt M, Dwojak T, Hoang H (2016) Is neural machine translation ready for deployment? A case study on 30 translation directions. In: Proceedings of the 9th international workshop on spoken language translation, Seattle, WA

Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. Prague Bull of Math Linguist 108:121–132

Klubička F, Toral A, Sánchez-Cartagena VM (2018) Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. Mach Translat, https://doi.org/10.1007/s10590-018-9214-x Accessed 27 Oct 2018

Koehn P, Knowles R (2017) Six challenges for neural machine translation. In: Proceedings of the first workshop on neural machine translation, pp 28–39

Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: 7th workshop on statistical machine translation. In: Proceedings of the workshop. Association for Computational Linguistics, Stroudsburg, pp 181–190

Koponen M, Salmi L (2017) Post-editing quality: analyzing the correctness and necessity of post-editor corrections. Linguist Antverp 16:137–148

Koponen M, Aziz A, Ramos L, Specia L (2012) Post-editing time as a measure of cognitive effort. In: O'Brien S, Simard M, Specia L (eds) (2012) Proceedings of the AMTA 2012 workshop on post-editing technology and practice. Association for machine translation in the Americas, pp 11–20

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. Springer, Heidelberg

Krings HP (2001) Repairing texts: Empirical investigations of machine translation post-editing process. The Kent State University Press, Kent

Kumpulainen M (2015) On the operationalisation of 'pauses' in translation process research. Int J Trans Interpret Res 7:47–58

Lacruz I, Shreve GM (2014) Pauses and Cognitive Effort in Post-editing. In: O'Brien S, Winther Balling S, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: processes and application. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 246–272

Lacruz I, Denkowski M, Lavie A (2014) Cognitive demand and cognitive effort in post-editing. In: Proceedings of the third workshop on post-editing technology and practice (WPTP-3), pp 73–84

Leal Fontes H (2013) Evaluating machine translation: preliminary findings from the first DGT-wide translators' survey. Lang Transl 6:10–11

Lörscher W (1986) Linguistic aspects of translation process: towards an analysis of translation performance. In: House J, Blum-Kulka S (eds) Interlingual and Intercultural Communication. Discourse and Cognition in Translation and Second Language Acquisition Studies. Gunter Narr Verlag, Tübingen, pp 277–292

Mossop B (2018) Evaluating the evaluators: quality assessment of revisers and revisions. Presentation at the Transius conference, Geneva, 18–20 June 2018

Mossop B (forthcoming) Appendix 3. Quantitative grading scheme. Appendix to appear in the new edition of revising and editing for translators

Mutta M (2016) Pausal Behavior in the Writing Processes of Foreign and Native Language Writers: The Importance of Defining the Individual Pause Length. In: Plane S, Bazerman C, Rondelli F, Donahue C, Applebee AN, Boré C, Carlino P, Marquilló Larruy M, Rogers P, Russell D (eds) Recherches en écriture: regards pluriels. Recherches Textuelles nro 13. Centre de Recherche sur les médiations (Crem): Université de Lorraine, pp 583–604

O'Brien S (2005) Methodologies for measuring the correlations between post-editing effort and machine translatability. Mach Transl 19:37–58

O'Brien S (2006) Pauses as indicators of cognitive effort in post-editing machine translation output. Across Lang Cult 7:1–21

Östling R, Scherrer Y, Tiedemann J, Tang G, Nieminen T (2017). The Helsinki neural machine translation system. In: Proceedings of the conference on machine translation (WMT), volume 2: Shared Task Papers, pp 338–347

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, pp 311–318. https://doi.org/10.3115/1073083.1073135

Peris Á, Casacuberta F (2018) Online learning for effort reduction in interactive neural machine translation. arXiv:1802.03594v1 [cs.CL]. Accessed 27 Oct 2018
Pirinen T (2008) Automatic finite state morphological analysis of finnish language using open source resources (in Finnish). Master's thesis, University of Helsinki
Pirinen T, Toral A, Rubino R (2016) Rule-based and statistical morph segments in English-to-finnish SMT. In: Proceedings of second international workshop on computational linguistics for uralic languages (IWCLUL), pp 60–73
Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. Prague Bull Math Linguist 93:7–16
Popović M (2018) Language-related issues for NMT and PBMT for English–German and English–Serbian. Mach Transl. https://doi.org/10.1007/s10590-018-9219-5. Accessed 27 Oct 2018
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: Proceedings of the 17th annual conference of the European Association for Machine Translation, EAMT 2014, pp 191–198
Rayner K, Pollatsek A, Ashby J, Clifton C Jr (2012) Psychology of reading, 2nd edn. Psychology Press, New York
Robin E (2018) A classification of revisional modifications. In: Horváth I (ed) Latest trends in hungarian translation studies. Budapest, OFFI-ELTE, pp 155–163
Sánchez-Cartagena VM, Toral A (2016) Abu-MaTran at WMT 2016 Translation task: deep learning, morphological segmentation and tuning on character sequences. Proceedings of the first conference on machine translation, Berlin, pp 362–370
Schaeffer M, Carl M, Lacruz I, Aizawa A (2016) Measuring cognitive translation effort with activity units. Baltic J Mod Comput 4:331–345
Shterionov D, Superbo R, Nagle P, Casanellas L, O'Dowd T, Way A (2018) Human versus automatic quality evaluation of NMT and PBSMT. Mach Transl. https://doi.org/10.1007/s10590-018-9220-z. Accessed 27 Oct 2018
Silfverberg M, Ruokolainen T, Lindén K, Kurimo M (2015) FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. Lang Resour Eval 50:863–878
Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. Proceedings of the 7th conference of the association for machine translation in the Americas. Association for machine translation in the Americas, pp 223–231
Snover M, Madnani N, Dorr B, Schwartz R (2009) Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In: Proceedings of the fourth workshop on statistical machine translation. Association for Computational Linguistics, Stroudsburg, pp 259–268
TAUS (2010) MT Post-editing guidelines. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines. Accessed 27 Oct 2018
Tiedemann J, Ginter F, Kanerva J (2015) Morphological segmentation and OPUS for finnish-English machine translation. In: Proceedings of the tenth workshop on statistical machine translation. Association for Computational Linguistics, Stroudsburg, pp 177–183
Tiedemann J, Cap F, Kanerva J, Ginter F, Stymne S, Östling R, Weller-Di Marco M (2016) Phrase-based SMT for finnish with more data, better models and alternative alignment and translation tools. In: Proceedings of the first conference on machine translation, Berlin, pp 391–398
Tirkkonen-Condit S, Jääskeläinen R (eds) (2000) Tapping and mapping the processes of translation and interpreting: outlooks on empirical research. John Benjamins, Amsterdam
Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th conference of the european chapter of the association for computational linguistics: volume 1, Long Papers, pp 1063–1073
Toral A, Wieling M, Way A (2018) Post-editing effort of a novel with statistical and neural machine translation. Front Digit Humanit 5:9
Toury G (2012) Descriptive translation studies and beyond. John Benjamins, Amsterdam
Vieira LN (2014) Indices of cognitive effort in machine translation post-editing. Mach Transl 28:187–216
Vieira LN (2017a) From process to product: links between post-editing effort and post-edited quality. In: Jakobsen AL, Mesa-Lao B (eds) Translation in transition: between cognition, computing and technology. John Benjamins, Amsterdam, pp 162–186
Vieira LN (2017b) Cognitive effort and different task foci in post-editing of machine translation: a think-aloud study. Across Lang Cult 18:79–105
Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser Ł, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian

G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016). Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144v2