

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25, 131-148.

<https://doi.org/10.26034/cm.jostrans.2016.303>

This article is published under a *Creative Commons Attribution 4.0 International* (CC BY):

<https://creativecommons.org/licenses/by/4.0>



© Maarit Koponen, 2016

Is machine translation post-editing worth the effort? A survey of research into post-editing and effort

Maarit Koponen, University of Helsinki

ABSTRACT

Advances in the field of machine translation have recently generated new interest in the use of this technology in various scenarios. This development raises questions over the roles of humans and machines as machine translation appears to be moving from the peripheries of the translation field closer to the centre. The situation which influences the work of professional translators most involves post-editing machine translation, i.e. the use of machine translations as raw versions to be edited by translators. Such practice is increasingly commonplace for many language pairs and domains and is likely to form an even larger part of the work of translators in the future. While recent studies indicate that post-editing high-quality machine translations can indeed increase productivity in terms of translation speed, editing poor machine translations can be an unproductive task. The question of effort involved in post-editing therefore remains a central issue. The objective of this article is to present an overview of the use of post-editing of machine translation as an increasingly central practice in the translation field. Based on a literature review, the article presents a view of current knowledge concerning post-editing productivity and effort. Findings related to specific source text features, as well as machine translation errors that appear to be connected with increased post-editing effort, are also discussed.

KEYWORDS

Machine translation, MT quality, MT post-editing, post-editing effort, productivity.

1. Introduction

In recent years, the translation industry has seen a growth in the amount of content being translated as well as pressure to increase the speed and productivity of translation. At the same time, technological advances in the field of machine translation (MT) development have led to wider availability of MT systems for various language pairs and improved MT quality. Combined, these factors have contributed to a renewed surge of interest in MT both in research and practice. This changing landscape of the translation industry raises questions on the roles of humans and machines in the field. For a long time, MT may have seemed relatively peripheral, with only limited use, but these recent advances are making MT more central to the translation field.

On the one hand, the availability of free online MT systems is enabling amateur translators as well as the general public to use MT. For example, the European Commission Directorate General for Translation (DGT) decided in 2010 to develop a new MT system intended not only as a tool for translators working at the DGT, but also for their customers for producing raw translations as self-service (Bonet 2013: 5). On the other hand, MT is also spreading in more professional contexts as it becomes integrated in widely used translation memory systems. While MT fully replacing human translators seems unlikely, human-machine interaction is certainly becoming a larger part of the work for many professional translators.

Despite improved MT quality in many language pairs, machine-translated texts are generally still far from publishable quality, except in some limited scenarios involving narrow domains, controlled languages and dedicated MT systems. Therefore, a common practice for including MT in the workflow is to use machine translations as raw versions to be further post-edited by human translators. This growing practical interest in the use of post-editing of MT is also reflected on the research side. Specific post-editing tools are being developed and the practices and processes involved are being studied by various researchers. Interest in post-editing MT is also reflected by a number of recent publications including an edited volume (O'Brien *et al.* 2014), a special journal issue (O'Brien and Simard 2014) and international research workshops focusing on the subject: the Workshops on Post-editing Technology and Practice that have been organised annually since 2012 are but one example. Information about the latest workshop in 2015 is available at <https://sites.google.com/site/wptp2015/>.

Studies have shown that post-editing high-quality MT can, indeed, increase the productivity of professional translators compared to manual translation 'from scratch' (see, for example, Guerberof 2009, Plitt and Masselot 2010). However, editing poor machine translation remains an unproductive task, as can be seen, for example, in the case of the less successful language pairs in the DGT trials (Leal Fontes 2013). For this reason, one of the key questions in post-editing MT is how much *effort* is involved in post-editing and the amount of effort that is acceptable: when is using MT really worth it? Estimating the actual effort involved in post-editing is important in terms of the working conditions of the people involved in these workflows. As Thicke (2013: 10) points out, the post-editors are ultimately the ones paying the price for poor MT quality. Objective ways of measuring effort are also important in the pricing of post-editing work. Pricing is often based on either hourly rates or determining a scale similar to the fuzzy match scores employed with translation memory systems (Guerberof Arenas 2010: 3). One problem with these practices is that measuring actual post-editing time is not

always simple and the number of changes needed may not be an accurate measure of the effort involved.

This paper presents a survey of research investigating the post-editing of MT and, in particular, the effort involved. Section 2 provides an overview of the history and more recent developments in the practical use of MT and post-editing. Section 3 presents a survey of the literature on post-editing in general, as well as a closer look into post-editing effort. Section 4 discusses the role of MT and post-editing as they evolve from a peripheral position towards a more central practice in the language industry.

2. Post-editing in practice – some history and current state

While the surge of interest in MT and post-editing workflows appears recent, it is in fact not a new issue; rather, post-editing is one of the earliest uses envisioned for MT systems. In his historical overview of post-editing and related research, García (2012: 293) notes that “[p]ostediting was a surprisingly hot topic in the late 1950s and early 1960s.” Edmundson and Hays (1958) outline one of the first plans for an MT system and post-editing process, which was intended for translating scientific texts from Russian into English at the RAND Corporation. In this early approach, it was assumed that the post-editor would work on the machine-translated text with a grammar code indicating the part of speech, case, number and other details of every word, and therefore would need “to have good command of English grammar and the technical vocabulary of the scientific articles being translated” (Edmundson and Hays 1958: 12). However, command of the source language was not required. According to García (2012: 294), this first documented post-editing program was discontinued in the early 1960s. In the 1960s, post-editing was used also by the US Air Force’s Foreign Technology Division and Euratom, but funding in the US ended in part because the report by the Automatic Language Processing Committee in 1966 (the so-called ALPAC report) found post-editing not to be worth the effort in terms of time, quality and difficulty compared to human translation (García 2012: 295-296).

While early uses of both MT as a technology and post-editing as a practice failed to live up to the initial visions, development continued. MT systems and post-editing processes were implemented in organisations such as the European Union and the Pan-American Health Organization as well as in some companies from the 1970s onwards (García 2012: 296-299). As studies have shown that MT can now produce sufficient quality to be usable in business contexts and increase the productivity of translators at least for some language pairs, post-editing workflows have become increasingly common (see, for example, Plitt and Masselot 2010, Zhechev 2014, Silva 2014). A recent study by Gaspari *et al.* (2015) gives an overview of various surveys carried out in recent years, in addition to

reporting their own survey of 438 stakeholders in the translation and localisation field. According to Gaspari *et al.* (2015: 13–17), 30% of their respondents were using MT and 21% considered it sure or at least likely that they would start using it in the future. Of those using MT, 38% answered that it was always post-edited, 30% never performed post-editing, whereas the remaining 32% used post-editing at varying frequencies. Overall, the surveys cited point to increasing use of MT and post-editing, increased demand for this service and expectations that the trend will continue.

As an example of the productivity gains often quoted, Robert (2013: 32) states that post-editing MT can increase the average number of words translated by a professional from 2,000 to 3,500 words per day. According to Guerberof Arenas (2010: 3), a productivity gain of 5,000 words per day is commonly quoted, but she points out that the actual figures will vary across different projects and post-editors. García (2011: 228) also notes that this figure involves certain conditions — professional, experienced post-editors, domain-specific MT systems and source text potentially pre-edited for MT — which may not always apply.

The use, and usability, of MT and post-editing also varies greatly in different language pairs. Discussing a survey of translators at the European Commission Directorate General for Translation, Leal Fontes (2013) presents the varying uptake rates for various language combinations. For example, translators working with the language pairs French-Spanish, French-Italian and French-Portuguese rated most MT segments reusable, but for English-German, English-Finnish and English-Estonian, translators considered MT only sufficient to suggest ideas for expressions or not usable at all (Leal Fontes 2013: 11). Some languages may present their own challenges: languages with rich morphology, such as Finnish, are known to be difficult for MT systems. Furthermore, language pairs with small markets often attract less interest in development work. On the user side, productivity studies at large companies like Autodesk have shown great variation in the gains achieved and amount of editing needed (Plitt and Masselot 2010, Zhechev 2014). Most of the practical use of post-editing involves professional contexts, but crowdsourcing has also been explored, for example, in disaster relief situations (Hu *et al.* 2011), websites (Tatsumi *et al.* 2012), and discussion forum contexts (Mitchell *et al.* 2013).

As post-editing workflows become increasingly common, there is a growing need for both organisation-specific and more general post-editing guidelines (see, for example, Guerberof 2009, 2010). A set of such guidelines by TAUS (2010), intended to help customers and language service providers in setting clear expectations and instructing post-editors, includes general recommendations for decreasing the amount of post-editing needed, as well as basic guidelines for carrying out post-editing to two defined quality levels. The first level of quality, termed 'good enough',

is defined as comprehensible and accurate so that it conveys the meaning of the source text, without necessarily being grammatically or stylistically perfect. At this level, the post-editor should make sure that the translation is semantically correct, that no information is accidentally added or omitted and that it does not contain offensive or inappropriate content. The second level would be 'publishable quality,' similar to that expected of human translation. In addition to being comprehensible and accurate, it should also be grammatically and stylistically appropriate. An International Standard (ISO/CD 18587:2014) is also being drafted with a view to specifying standardised requirements for the post-editing process and post-editor competences. This draft standard also specifies two post-editing levels (light and full post-editing), similar to the two levels of the TAUS guidelines (ISO/CD 18587: 6). An earlier definition with three levels (rapid, minimal and full post-editing) can be found in Allen (2003).

3. Post-editing research

The first study focusing on post-editing can be dated to 1967 (García 2012: 301). This study, conducted by Orr and Small (1967), compared reading comprehension by test subjects who read a raw machine translation, a post-edited machine translation or a manual translation of scientific texts. The test measured the percentage of correct answers, time taken to finish the comprehension test and efficiency, which was measured as the number of correct answers per 10-minute time interval. While they found some differences between post-edited and manually translated texts, their results indicated that differences were much larger between reading machine-translated and post-edited versions (Orr and Small 1967: 9). For a more comprehensive historical perspective on post-editing studies see García (2012). More recently, research both on the academic and the commercial sides has focused on productivity: post-editing speed or number of pages post-edited in a given timeframe compared to translation from scratch, often combined with some form of quality evaluation. Post-editing effort, which can be defined in different ways (see Section 3.4), has also received much interest lately.

The following sub-sections present an overview of post-editing studies. Sub-section 3.1 discusses productivity studies and sub-section 3.2 presents studies related to the quality of the post-edited texts. Sub-section 3.3 presents studies investigating the viability of post-editing without access to the source text, i.e. so-called monolingual post-editing. Sub-section 3.4 addresses the concept of post-editing effort, ways of measuring it and findings related to features potentially affecting the amount of effort.

3.1. Productivity

Extensive productivity studies for post-editing MT compared to manual translation have been carried out in various companies. A study at

Autodesk, reported by Plitt and Masselot (2010), involved twelve professional translators working from English into French, Italian, German or Spanish and compared their throughput (words translated per hour) when post-editing MT to manual translation. On the average, post-editing was found to increase throughput by 74%, although there was considerable variation between different translators (Plitt and Masselot 2010: 10). The continuation study reported by Zhechev (2014) included further target languages and showed large variation in productivity gains between different language pairs.

By contrast, studies by Carl *et al.* (2011) and García (2010) did not find significant improvements in productivity when comparing post-editing and manual translation. However, Carl *et al.* (2011: 137-138) acknowledge that their test subjects did not have experience in post-editing or in the use of the tools involved, and the participants in García's study (2010: 7) were trainee translators, which may have some connection to the lack of productivity gains. While data involving relatively inexperienced subjects may not entirely reflect the processes and productivity of experienced, professional post-editors, it should be noted that Guerberof Arenas (2014) found no significant differences in processing speed when comparing experienced versus novice post-editors.

In general, studies have shown considerable variation between editors in the speed of post-editing (Plitt and Masselot 2010, Sousa *et al.* 2011) and post-editors appear to differ more in terms of post-editing time and number of keystrokes than in terms of number of textual changes made (Tatsumi and Roturier 2010, Koponen *et al.* 2012). These studies have also found varying editor profiles: speed and number of keystrokes or the number of edits made do not necessarily correlate and post-editors have different approaches to the post-editing process.

In summary, research into the productivity of post-editing shows partly conflicting results: some studies have reported quite notable productivity gains when comparing post-editing to manual translation, while others have not found significant differences. The productivity gains appear to depend, at least partly, on specific conditions. These conditions relate to sufficiently high-quality machine translation, which is currently achievable for certain language pairs, and machine translation systems geared toward the specific text type being translated. The post-editors' familiarity with the tools and processes involved in the post-editing task may also play a role. In addition, comparison of the post-editors themselves shows individual variation: the differing work processes of different people appear to affect how much they benefit from the use of machine translation.

3.2. Quality

Studies have also addressed quality, as there have been concerns that the quality of post-edited texts would be lower than that of manually translated texts. Fiederer and O'Brien (2009) addressed this question in an evaluation experiment that compared a set of sentences translated either manually or by post-editing MT. The different versions were rated separately for clarity (intelligibility of the sentence), accuracy (same meaning as the source sentence) and style (appropriateness for purpose, naturalness and idiomaticity). Fiederer and O'Brien (2009: 62-63) found that post-edited translations were rated higher for clarity (although only very slightly) and accuracy, but lower in terms of style. When asked to select their favourite translated version of each sentence, the evaluators mostly preferred the manually translated versions, which may indicate that they were most concerned with style. Using a slightly different approach, Carl *et al.* (2011) had evaluators rank manually translated and post-edited versions of sentences in order of preference, and found a slight, although not significant, preference for the post-edited translations.

Some studies have approached quality in terms of the number of errors found in the translations. In Plitt and Masselot's (2010) study, both the manually translated and post-edited texts were assessed by the company's quality assurance team according to the criteria applied to all their publications. All translations, whether manually translated or post-edited, were deemed acceptable for publication, but, somewhat surprisingly, the evaluators in fact marked more manually translated sentences as needing corrections (Plitt and Masselot 2010: 14-15). In another error-based study, García (2010) used the guidelines by the Australian National Accreditation Authority for Translators and Interpreters to compare manual and post-edited translations. Post-edited MT and manually translated texts received comparable evaluations, with post-edited texts in fact receiving slightly higher average marks from both of the evaluators involved in the study (García 2010: 13-15).

Based on the studies assessing the quality of post-edited texts, it appears that post-editing can lead to quality levels similar to manually translated texts. In fact, depending on the quality aspect examined, post-edited texts are sometimes even evaluated as better.

3.3. Monolingual post-editing

While post-editing practice and research mainly involve bilingual post-editors correcting the MT based on the source text, some studies have also investigated scenarios where a post-editor would work on the MT without access to the source text. In these scenarios, the central question becomes whether monolingual readers are able to understand the source meaning based on a machine translation alone. Koehn (2010) describes a set-up where test subjects post-edited MT without access to the source

text. The correctness of the post-edited sentences was then evaluated based on a simple standard of correct or incorrect, where a correct sentence was defined as “a fluent translation that contains the same meaning in the document context” (Koehn 2010: 541). Depending on the language pair and MT system, between 26% and 35% of the sentences in Koehn’s study were accepted as correct. A similar approach was utilised by Callison-Burch *et al.* (2010), whose results show great variation in the percentages of sentences accepted as correct, depending on the language pair and system used. For the best system in each language pair investigated, the rates of sentences evaluated as correct ranged from 54% to 80%, whereas for the lowest-rated systems, in some cases less than 10% of the sentences were accepted as correct after post-editing (Callison-Burch *et al.* 2010: 28). One difficulty with the evaluation against this standard is that it is not clear whether sentences were rejected due to errors in language or in meaning.

Other monolingual post-editing studies have utilised separate evaluation for language and meaning. In their study of monolingual post-editing of text messages for emergency responders in a catastrophe situation, Hu *et al.* (2011) carried out an evaluation on separate five-point scales for fluency of language and adequacy of meaning; 24% to 39% of sentences (depending on system and test set) achieved the highest adequacy score as rated by two evaluators. For language, the percentage of sentences achieving the highest fluency score ranged from under 10% to over 30% (Hu *et al.* 2011: 401-403). A study carried out by Koponen and Salmi (2015) also investigated the question of language and content errors in a monolingual post-editing experiment. The post-edited versions were evaluated separately for correct language and correct meaning, and the results show that test subjects were able to correct about 30% of the sentences for both language and meaning, with an additional 20% corrected for meaning but still containing language errors. The results reported in Koponen and Salmi (2015: 131) suggest that many of the remaining language errors, such as punctuation or spelling errors, appear to be due to inattentiveness on the part of at least some test subjects. In some cases, the corrections themselves contained errors ranging from misspellings to grammar issues caused by correcting only part of the sentence.

The success of monolingual post-editing has also been directly compared to bilingual post-editing, where the post-editors have access to the source text. Mitchell *et al.* (2013) carried out such a comparison in their study of machine-translated online user forum posts and evaluated quality with regard to fluency, comprehensibility and fidelity of meaning by assessing whether the post-editing improved or possibly lowered the quality scores. The results showed variation in the different quality measures. Monolingual post-editing was found to improve the fidelity scores less than bilingual post-editing (43% vs 56% of sentences) in the English-German language pair, but slightly more often (67% vs 64% of sentences) in the

English-French language pair. For improving comprehensibility, monolingual post-editing was in both cases less successful than bilingual: in the English-German language pair, 57% of sentences improved in monolingual post-editing compared to 64% of sentences in the bilingual condition, a range that was smaller than in the English-French pair, where monolingual post-editing was found to improve only 48% of sentences, compared to 63% of sentences improved by bilingual post-editing. In terms of fluency, monolingual and bilingual post-editing were equally successful in the English-French pair (63% of sentences improved in both conditions), and very similar in the English-German pair (67% vs 70% of sentences).

Although early researchers envisioning post-editing of MT, like Edmundson and Hays (1958), assumed only monolingual post-editing would be necessary, these later studies do not support its feasibility. Even if monolingual post-editors are able to improve the language, at least so far MT quality has not been sufficient to reliably convey the meaning of the source text. In many cases, it even seems that the post-editors are not very attentive to language errors either.

3.4. Post-editing effort

While much of the interest, particularly on the industry side, focuses on the speed or number of pages produced through post-edited machine translation compared to manual translation, this aspect reflects only part of the effort involved. In one of the most comprehensive early studies on post-editing, Krings (2001: 178-182) distinguishes three separate but interlinked dimensions of post-editing effort: temporal, technical, and cognitive. The temporal aspect of post-editing is formed by the technical operations (insertions, deletions and reordering of words) necessary for corrections, as well as the cognitive effort involved in detecting the errors and planning the corrections. When attempting to measure the effort involved in post-editing and to determine the features that potentially lead to increased effort, scholars have used different approaches.

One possible way of investigating post-editing effort is, naturally, to ask the people involved to assess how much effort they perceive or expect during the post-editing of a given text. With the intention of creating a specific post-editing effort evaluation method, Specia *et al.* (2010: 43) propose a four-point scale to be used by professional translators for rating machine-translated sentences to indicate the need for post-editing. A slightly different scale was used by Callison-Burch *et al.* (2012), who asked editors, after post-editing, to evaluate on a five-point scale how much of the sentence needed to be edited. The results of such evaluations obviously vary greatly depending on the quality of the MT. For example, in the data used by Specia *et al.* (2010: 43) average scores for 4,000 sentences translated from English to Spanish ranged from 1.3 to 2.8 on a scale from 1 (full retranslation necessary) to 4 (no post-editing needed).

Such scores given by human evaluators are often considered complicated (see, for example, Callison-Burch *et al.* 2012: 25) because they are to some extent subjective, and different people may have different perceptions of the effort involved. A study by Sousa *et al.* (2011) has, however, shown that sentences requiring less time to edit are more often tagged as low effort by evaluators. The human scores may therefore be useful in providing a rough idea of whether the output of a given machine translation system is suitable for post-editing.

In many cases, the determination of effort is based on post-editing time. Although time can be considered the most visible part of post-editing effort (Krings 2001: 179), collecting accurate information about post-editing time is not necessarily easy in practical work contexts. For example, post-editors' self-reports of the time used may not be complete or sufficiently detailed, and recording accurate information would require specialised tools that are not always available (for a more detailed discussion, see Moran *et al.* 2014). Interest in collecting data for investigating post-editing effort has led to the development of various tools (see, for example, Aziz *et al.* 2012, Elming *et al.* 2014, Moran *et al.* 2014), which include different functionalities to record keylogging and sometimes eye tracking data in addition to time.

Technical effort can be measured using computerised metrics, such as the commonly used Human-targeted Translation Edit Rate or HTER (Snover *et al.* 2006). These metrics compare how many words have been changed (added, deleted, replaced or moved) between the machine-translated version and the post-edited version of a given sentence, and thereby reflect the technical effort to some extent. However, as Elming *et al.* (2014) point out, the use of keylogging to track how the changes were made is a better measure of the actual technical effort. Furthermore, the number of changes does not always correspond to the time needed for post-editing. For example, Tatsumi (2009) found that post-editing time is sometimes longer or shorter than expected based on the number of edits.

Much attention has recently been paid to which factors influence post-editing effort. Tatsumi (2009) suggested source sentence length and structure, as well as specific types of errors, as possible explanations for cases where post-editing takes longer than expected based on the number of edits. Studies have since explored what specific types of errors in the machine-translated output or specific source text characteristics might be linked to increased effort. Temnikova (2010) examined different types of edits in relation to post-editing time and found that machine translations of texts that had been simplified using controlled language rules were faster to edit; they contained more edits assumed to be simple, such as changing the morphological form of a word, and less edits considered to be more cognitively demanding, such as correcting the word order or editing mistranslated idioms (Temnikova 2010: 3488-3489). Other studies comparing sentences with human evaluation scores or post-editing times

have also found that sentences involving less effort, as indicated by higher human scores or shorter post-editing times, tend to contain more edits related to word forms and simple substitutions of words of the same word class, while sentences with low scores or long post-editing times involve more edits related to word order, edits where the word class was changed, and corrections of mistranslated idioms (Koponen 2012, Koponen *et al.* 2012).

In studies concerned with the effect of source text features and their relation to post-editing effort, sentence length, in particular, has been found to have some effect. Both very long and very short sentences have been found to have longer post-editing times than expected based on the number of edits (Tatsumi 2009, Tatsumi and Roturier 2010), and very long sentences tend to score low in human evaluations (Koponen 2012). Another factor may be sentence structure. Incomplete sentences as well as complex and compound sentences were found to be associated with longer post-editing times by Tatsumi and Roturier (2010: 47-49). In a study investigating specific source text patterns within sentences, Aziz *et al.* (2014) found that there appears to be some connection between longer post-editing times and edits related to passages containing specific types of verbs or noun sequences. Vieira (2014), on the other hand, found some connection between increased cognitive effort and prepositional phrases, as well as instances where repetitions of the same words appear within a sentence.

Overall, cognitive effort appears difficult to capture. In an early approach, Krings (2001) utilised think-aloud protocols (TAP), where the post-editors reported their actions verbally during the post-editing process. While this approach has been widely used in translation process research in the past, it has many drawbacks, such as slowing down the process and changing the cognitive processing involved (for a full discussion, see O'Brien 2005: 41-43). Pauses found in keylogging data recorded during post-editing have attracted attention as a measure of cognitive effort. O'Brien (2005) found some correlation between potentially difficult source text features identified by comparing different post-editors' versions and pauses recorded in keyboard logging. In a further study, O'Brien (2006) analysed pauses during the post-editing of sentences with source text characteristics previously shown to involve increased effort and sentences without these features, but found difficulties in connecting the location and duration of pauses to the cognitive processing. While these studies and much of the earlier research involves identifying long pauses, a different approach which examines clusters of short pauses has been explored by Lacruz *et al.* (2012) and Lacruz and Shreve (2014). The rationale, as described by Lacruz and Shreve (2014: 263), is that segments requiring more cognitive post-editing effort would contain a higher density of short pauses than segments requiring little effort. They identify edits related to incorrect syntax and mistranslated idioms as some

of the instances where increased cognitive effort is required (Lacruz and Shreve 2014: 267).

In summary, the effort involved in post-editing can be investigated in terms of post-editing time, the technical effort of performing the corrections and the effort perceived by humans. Studies related to these aspects have so far identified some features which appear connected to increased effort. They may relate to source text characteristics, such as the length and structure of the sentences, or specific features like noun sequences. Others relate to errors in the machine-translated output, leading to edits involving word order and correction of mistranslated idioms, for example. Questions still remain in determining the amount of effort involved in post-editing, particularly the amount of cognitive effort. New technology like eye tracking, which is based on the assumption that instances where the gaze fixates on a word for a long duration or multiple times indicate increased cognitive effort (see Carl *et al.* 2011: 138), has been used in recent studies (e.g. Vieira 2014). It may, in the future, offer further information about the features affecting effort.

4. MT and post-editing – from the peripheries towards the centre

The title of this paper posed the question of whether MT and post-editing are worth the effort involved. To investigate this question, an overview was presented of studies related to various aspects of the use of MT, post-editing and the effort involved. Based on this review, the answer appears to be: yes, when used in suitable conditions. The overall pattern arising from the literature seems to be that post-editing machine translation can increase the productivity (Section 3.1) of translators in terms of speed, while retaining or in some cases even improving the quality of their translations (Section 3.2). However, such benefits are not always guaranteed. Rather, they are dependent on the quality of the machine translations, as can be seen from the widely varying results for different machine translation systems, different texts and language pairs. There also appear to be other factors involved. For example, individual translators' work processes and experience may affect whether, and to what extent, they benefit from the machine translation as a raw version. In addition, post-editing does not currently appear feasible if the post-editors have no access to the source text (Section 3.3), although this has been suggested as a potential scenario. Research has also identified some potential features affecting the amount of effort required in post-editing (Section 3.4), but questions still remain as to how to accurately determine effort.

The overview of the use of machine translation, post-editing and research on these issues also shows how post-editing machine-translated texts is increasingly becoming a part of the translation workflow, simultaneously prompting new research interests. In this way, MT and post-editing are moving from a rather peripheral position to a more central place in the

translation field in different ways: as a practice within the translation industry, a research topic for Translation Studies and a topic for translator training. This move to a more central position is perhaps most evident in professional practice. As seen in Section 2, post-editing of MT is indeed already an established part of the translation workflow in many professional contexts. While MT quality, and therefore also the feasibility of post-editing, still varies in different language pairs, with efforts by organisations such as the European Commission as well as language service providers, the development seems likely to continue. Research into MT and post-editing has for a long time focused mainly on the technical side of MT system development, while in the field of Translation Studies, MT has been a rather peripheral issue. The situation has, however, been changing, with, for example, Fiederer and O'Brien (2009: 52) arguing for the need to engage researchers in the Translation Studies community more actively in the MT field. As human-machine interaction has increased in professional practice, post-editing processes and effort in particular have, indeed, recently become more central areas of interest to translation scholars.

The increased practical use suggests that at least in some contexts machine translation and post-editing certainly has been found to be worthwhile. At the same time, research has also provided evidence that post-editing machine-translated output can increase productivity in situations suited to the post-editing process and in this manner be worth the effort. However, post-editing does not appear to be suited for all scenarios. One of the key questions that remains is how to determine *when* post-editing is, or is not, worth the effort. Accurate measurement of the actual effort would be important as it has implications not only for productivity, but also for the working conditions of the post-editors.

The increased use of machine translation and post-editing workflows has already changed the role of humans and machines in the field, and will likely continue to do so. Some concerns have also been expressed that the technology might even be pushing humans out of the field. While machines alone might be suitable for some specific purposes, such as situations where the type of text to be translated is strictly limited, or situations where an imperfect translation conveying the basic content suffices, machine translation fully replacing the human translator appears unlikely based on the current knowledge of the potential of machine translation. Of course, predicting future developments of new technology or practices is difficult, but among the many varied translation scenarios, many still remain where the human is essential and the machine is only a potential tool.

Bibliography

- **Allen, Jeff** (2003). "Post-editing." Harold Somers (ed.) (2003) *Computers and Translation. A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins, 297–317.

- **Aziz, Wilker, Sheila C. M. Sousa and Lucia Specia** (2012). "PET: A Tool for Post-editing and Assessing Machine Translation." Nicoletta Calzolari *et al.* (eds) (2012) *Proceedings of the 8th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 3982–3987.
- **Aziz, Wilker, Maarit Koponen and Lucia Specia** (2014). "Sub-sentence Level Analysis of Machine Translation Post-editing Effort." O'Brien *et al.* (2014), 170–199.
- **Bonet, Josep** (2013). "No rage against the machine." *Languages and Translation* 6, 4–5.
http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue_06_en.pdf (consulted 11.6.2015).
- **Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar F. Zaidan** (2010). "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation." Chris Callison-Burch *et al.* (eds) (2010) *ACL 2010: Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Proceedings of the workshop*. Stroudsburg, PA: Association for Computational Linguistics, 17–53.
- **Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia** (2012). Chris Callison-Burch *et al.* (eds) (2012) "Findings of the 2012 workshop on statistical machine translation." *7th Workshop on Statistical Machine Translation. Proceedings of the workshop*. Stroudsburg, PA: Association for Computational Linguistics, 10–51.
- **Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt Lykke Jakobsen** (2011). "The process of post-editing: A pilot study." Bernadette Sharp *et al.* (eds) (2011) *Proceedings of the 8th International NLPSC workshop. Special theme: Human-machine interaction in translation*. Copenhagen Studies in Language 41. Frederiksberg: Samfundslitteratur, 131–142.
- **Edmundson, Harold P. and David G. Hays** (1958). "Research methodology for machine translation." *Mechanical Translation* 5(1), 8–15.
- **Elming, Jakob, Laura Winther Balling and Michael Carl** (2014). "Investigating User Behaviour in Post-Editing and Translation using the CASMACAT Workbench." O'Brien *et al.* (2014), 147–169.
- **Fiederer, Rebecca and Sharon O'Brien** (2009). "Quality and machine translation: A realistic objective?" *The Journal of Specialised Translation* 11, 52–74.
http://www.jostrans.org/issue11/art_fiederer_obrien.pdf (consulted 11.6.2015).
- **García, Ignacio** (2010). "Is machine translation ready yet?" *Target* 22(1), 7–21.
- — (2011). "Translating by post-editing: Is it the way forward?" *Machine Translation* 25(3), 217–237.
- — (2012). "A brief history of postediting and of research on postediting." Anthony Pym and Alexandra Assis Rosa (eds) (2012) *New Directions in Translation Studies*. Special Issue of *Anglo Saxonica* 3(3), 292–310.
- **Gaspari, Federico, Hala Almaghout and Stephen Doherty** (2015). "A survey of machine translation competences: insights for translation technology educators and practitioners." *Perspectives: Studies in Translatology*, 1–26.
<http://dx.doi.org/10.1080/0907676X.2014.979842> (consulted 11.6.2015).

- **Guerberof, Ana** (2009). "Productivity and quality in MT post-editing." Marie-Josée Goulet *et al.* (eds) (2009) *Beyond Translation Memories Workshop. MT Summit XII, Ottawa*. Association for Machine Translation in the Americas.
- **Guerberof Arenas, Ana** (2010). "Project management and machine translation." *Multilingual* 21(3), 1–4.
- — (2014). "The role of professional experience in post-editing from a quality and productivity perspective." O'Brien *et al.* (2014), 51–76.
- **Hu, Chang, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek and Benjamin B. Bederson** (2011). "The value of monolingual crowdsourcing in a real-world translation scenario: simulation using Haitian Creole emergency SMS messages." Chris Callison-Burch *et al.* (eds) (2011) *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT '11)*. Stroudsburg PA: Association for Computational Linguistics, 399–404.
- **ISO/CD 18587:2014** (2014). *Translation Services – Post-editing of machine translation output – Requirements*. International Organization for Standardization.
- **Koehn, Philipp** (2010). "Enabling monolingual translators: Post-editing vs. options." Ron Kaplan *et al.* (eds) (2010) *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg PA: Association for Computational Linguistics, 537–545.
- **Koponen, Maarit** (2012). "Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations." Chris Callison-Burch *et al.* (eds) (2012) *7th Workshop on Statistical Machine Translation. Proceedings of the workshop*. Stroudsburg, PA: Association for Computational Linguistics, 181–190.
- **Koponen, Maarit, Wilker Aziz, Luciana Ramos and Lucia Specia** (2012). "Post-Editing Time as a Measure of Cognitive Effort." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012) *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*. Association for Machine Translation in the Americas, 11–20.
- **Koponen, Maarit and Leena Salmi** (2015). "On the correctness of machine translation: A machine translation post-editing task". *The Journal of Specialised Translation* 23, 118–136. http://www.jostrans.org/issue23/art_koponen.pdf (consulted 11.6.2015).
- **Krings, Hans P.** (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (ed. Geoffrey S. Koby). Kent, Ohio: Kent State University Press.
- **Lacruz, Isabel, Gregory M. Shreve and Erik Angelone** (2012). "Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012) *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*. Association for Machine Translation in the Americas, 21–30.
- **Lacruz, Isabel and Gregory M. Shreve** (2014). "Pauses and Cognitive Effort in Post-editing." O'Brien *et al.* (2014), 246–272.
- **Leal Fontes, Hilário** (2013). "Evaluating Machine Translation: preliminary findings from the first DGT-wide translators' survey." *Languages and Translation* 6, 10–11.

http://ec.europa.eu/dgs/translation/publications/magazines/languagetranslation/documents/issue_06_en.pdf (consulted 4.5.2015).

- **Mitchell, Linda, Johann Roturier and Sharon O'Brien** (2013). "Community-based post-editing of machine-translated content: monolingual vs. bilingual." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2013) *Workshop Proceedings: Workshop on Post-editing Technology and Practice (WPTP-2)*. Allschwil: The European Association for Machine Translation, 35–44.
- **Moran, John, David Lewis and Christian Saam** (2014). "Analysis of Post-editing Data: A Productivity Field Test Using an Instrumented CAT Tool." O'Brien *et al.* (2014), 126–146.
- **O'Brien, Sharon** (2005). "Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability." *Machine Translation* 19(1), 37–58.
- — (2006). "Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output." *Across Languages and Cultures* 7, 1–21.
- **O'Brien, Sharon, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia** (eds) (2014). *Post-editing of Machine Translation: Processes and Applications*. Newcastle: Cambridge Scholars Publishing.
- **O'Brien, Sharon and Michel Simard** (2014). "Introduction to special issue on post-editing." *Machine Translation* 28(3), 159–164.
- **Orr, David B. and Victor H. Small** (1967). "Comprehensibility of Machine-aided Translations of Russian Scientific Documents." *Mechanical Translation and Computational Linguistics* 10(1-2), 1–10.
- **Plitt, Mirko and François Masselot** (2010). "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context." *The Prague Bulletin of Mathematical Linguistics* 93, 7–16. <http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf> (consulted 11.6.2015).
- **Robert, Anne-Marie** (2013). "Vous avez dit post-éditrice ? Quelques éléments d'un parcours personnel." *The Journal of Specialised Translation* 19, 29–40. http://www.jostrans.org/issue19/art_robert.pdf (consulted 11.6.2015).
- **Silva, Roberto** (2014). "Integrating Post-editing MT in a Professional Translation Workflow." O'Brien *et al.* (2014), 24–50.
- **Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul** (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." Laurie Gerber *et al.* (eds) (2006) *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, 223–231.
- **Sousa, Sheila C. M., Wilker Aziz and Lucia Specia** (2011). "Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD subtitles." Galia Angelova *et al.* (eds) (2011) *Proceedings of the Recent Advances in Natural Language Processing Conference*. Hissar, Bulgaria: RANLP 2011 Organising Committee, 97–103.
- **Specia, Lucia, Dhvaj Raj and Marco Turchi** (2010). "Machine Translation Evaluation Versus Quality Estimation." *Machine Translation* 24(1), 39–50.

- **Tatsumi, Midori** (2009). "Correlation Between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors." Laurie Gerber et al. (eds) (2009) *Proceedings of the MT Summit XI*. Association for Machine Translation in the Americas, 332–339.
- **Tatsumi, Midori and Johann Roturier** (2010). "Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship?" Ventsislav Zhechev (ed.) (2010) *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10)*. Association for Machine Translation in the Americas, 43–51.
- **Tatsumi, Midori, Takako Aikawa, Kentaro Yamamoto and Hitoshi Isahara** (2012). "How Good Is Crowd Post-Editing? Its Potential and Limitations." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012) *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*. Association for Machine Translation in the Americas, 69–77.
- **TAUS** (2010). *Machine Translation Post-editing Guidelines*. <https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines> (consulted 11.6.2015)
- **Temnikova, Irina** (2010). "A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment." Nicoletta Calzolari et al. (eds) (2010) *Proceedings of the 7th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 3485–3490.
- **Thicke, Lori** (2013). "The industrial process for quality machine translation." *The Journal of Specialised Translation* 19, 8–18. http://www.jostrans.org/issue19/art_thicke.pdf (consulted 11.6.2015)
- **Vieira, Lucas Nunes** (2014). "Indices of cognitive effort in machine translation post-editing." *Machine Translation* 28(3), 187–216.
- **Zhechev, Ventsislav** (2014). "Analysing the Post-Editing of Machine Translation at Autodesk." O'Brien et al. (2014), 2–13.

Biography

Maarit Koponen has an MA in English Philology (2005) from the University of Helsinki. She is currently writing her PhD in Language Technology (focusing on machine translation evaluation and post-editing) at the University of Helsinki in the Department of Modern Languages. She has taught computer-assisted translation and post-editing at the University of Helsinki and professional translation at the University of Eastern Finland. She has also worked as a professional translator for over five years.

E-mail: maarit.koponen@helsinki.fi

