Department of Modern Languages
Faculty of Arts
University of Helsinki

Machine Translation Post-editing and Effort

Empirical Studies on the Post-editing Process

Maarit Koponen

ACADEMIC DISSERTATION
to be publicly discussed, by due permission of the Faculty of Arts
at the University of Helsinki, in Auditorium XII, Main Building,
on the 19th of March, 2016 at 10 o'clock.

Helsinki 2016

Supervisor
Prof. Lauri Carlson, University of Helsinki, Finland

Pre-examiners
Dr. Sharon O'Brien, Dublin City University, Ireland
Dr. Jukka Mäkisalo, University of Eastern Finland, Finland

Opponent
Dr. Sharon O'Brien, Dublin City University, Ireland

ISBN 978-951-51-1974-2 (paperback)
ISBN 978-951-51-1975-9 (PDF)
Unigrafia
Helsinki 2016

Abstract

This dissertation investigates the practice of machine translation post-editing and the various aspects of effort involved in post-editing work. Through analyses of edits made by post-editors, the work described here examines three main questions: 1) what types of machine translation errors or source text features cause particular effort in post-editing, 2) what types of errors can or cannot be corrected without the help of the source text, and 3) how different indicators of effort vary between different post-editors.

The dissertation consists of six previously published articles, and an introductory summary. Five of the articles report original research, and involve analyses of post-editing data to examine questions related to post-editing effort as well as differences between post-editors. The sixth article is a survey presenting an overview of the research literature. The research reported is based on multiple datasets consisting of machine translations and their post-edited versions, as well as process and evaluation data related to post-editing effort. The dissertation presents a mixed methods study combining qualitative and quantitative approaches, as well as theoretical and analytical tools from the fields of language technology and translation studies. Data on edits performed by post-editors, post-editing time, keylogging data, and subjective evaluations of effort are combined with error analyses of the machine translations in question, and compared for various post-editors.

The results of this dissertation provide evidence that, in addition to the number of edits performed, post-editing effort is affected by the type of edits as well as source text features. Secondly, the results show that while certain language errors can be corrected even without access to the source text, certain other types that more severely affect the meaning cannot. Thirdly, the results show that post-editors' speed and the amount of editing they perform differ, and that various profiles can be identified in terms of how the edits are planned and carried out by the post-editors. The results of this work may have both theoretical and practical implications for the measurement and estimation of post-editing effort.

Acknowledgements

From the first drafting of ideas for a PhD topic to finally completing my dissertation, this project has been a long and sometimes winding road. Along the way, I have been fortunate to have the support and guidance of many people to whom I wish to express my heartfelt gratitude. First and foremost, I am indebted to my supervisor, Professor Lauri Carlson, for all his invaluable support, encouragement, insight, and direction. From the very first steps on this road to where I stand today, your guidance has always helped me find the way forward.

During the dissertation project, I have benefited greatly from working with other researchers whose expertise and perspectives taught me much and helped shape this dissertation. For these rewarding collaborations, and sharing of ideas and inspiration, I thank my co-authors, Professor Leena Salmi, Dr Lucia Specia, Dr Wilker Aziz, and Luciana Ramos.

I also wish to thank Dr Sharon O'Brien and Dr Jukka Mäkisalo for their constructive and encouraging comments as the pre-examiners of this dissertation, as well as for valuable suggestions and comments on papers and presentations already in the earlier stages of this project. My thanks also to all of the anonymous reviewers of the original publications for their feedback, as well as the editors of these publications and revisers of both the original papers and this summary who have helped improve my writing.

Along the way, many people have helped me find my place in the academic community, showed interest in my work and opened doors to new opportunities. I am especially grateful to Liisa Tiittula and Kristiina Taivalkoski-Shilov for so many valuable opportunities and all your help. For all their support, questions, perspectives, and suggestions, I wish to thank my seniors and peers at the University of Helsinki, particularly Pirjo Kukkonen, Tuija Kinnunen, Olli Philippe Lautenbacher, Kimmo Koskenniemi, Krister Lindén, as well as my seniors and peers in the Langnet Doctoral Programme, particularly Kaisa Koskinen, Nina Pilke, Merja Koskela and other supervisors and students in the *Multilingualism and Professional Communication* and *Language Technology* sub-programs. I am also grateful to Professor Riitta Jääskeläinen, Esa Penttilä and others at the University of Eastern Finland for the opportunity to work there and for their assistance. For inspiring discussions on translation quality and for the opportunity to carry out practical work in the field, my thanks to Aarne Ranta, Jussi Rautio, Seppo Nyrkkö, and others in the EU Molto project.

For hours of inspiring discussions and invaluable peer support, I sincerely thank all my fellow PhD students who have celebrated and commiserated with me along the way – especially the "Monday study circle": Meri Päivärinne, Minna Hjort, Mika Loponen, Maija Hirvonen, Juha Eskelinen, Léa Huotari and others.

My work has been financially supported at various stages and in various forms by the Langnet Doctoral Programme and the Translation Studies and Terminology Research Group at the Department of Modern Languages, University of Helsinki. I am most grateful for all the financial and practical assistance.

This dissertation project would never have been possible without my family and friends. Particularly I am forever grateful to my parents Leena and Markku for encouraging my passion for knowledge all my life, and for all their support in countless ways. I thank all my family and friends for being there for me, and particularly Kati, for making sure I never take myself too seriously.

Finally, and above all, my deepest thanks and appreciation to Juha for absolutely everything. Thank you for being by my side every step of the way.

Helsinki, February 23, 2016
Maarit Koponen

List of Original Articles

This dissertation consists of an introduction and the following peer-reviewed publications (in chronological order). These publications are reproduced at the end of the print version of the thesis.

- Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190.
- Maarit Koponen, Wilker Aziz, Luciana Ramos and Lucia Specia. 2012. Post-editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*, pages 11–20.
- Maarit Koponen. 2013. This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 1–9.
- Wilker Aziz, Maarit Koponen and Lucia Specia. 2014. Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, and L. Specia (eds). *Post-editing of Machine Translation: Processes and Applications*, pages 170–199. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Maarit Koponen and Leena Salmi. 2015. On the correctness of machine translation: A machine translation post-editing task. In *Journal of Specialised Translation* 23, pages 118–136.
- Maarit Koponen. 2016. Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort. In *Journal of Specialised Translation* 25, pages 131–148.

Acronyms

HTER	human-targeted translation edit rate
MT	machine translation
PE	post-editing
PEMT	post-editing of machine translation
POS	part-of-speech
SL	source language
ST	source text
TL	target language
TT	target text

Contents

Chapter 1
Introduction

In recent years, the translation industry has seen a growth in the amount of content to be translated as well as pressure to increase the speed and productivity of translation. At the same time, technological advances in the development of machine translation systems have led to machine translation finding its place in professional contexts. The scenario most influencing the work of professional translators involves the use of machine translations as raw versions to be post-edited. This practice – generally termed *post-editing of machine translation* (PEMT) or simply *post-editing* (PE) – 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.

The purpose of this academic dissertation is to investigate the practice of machine translation post-editing and the various aspects of effort involved in post-editing work. Through analyses of edits made by post-editors and indicators of effort like post-editing time, amount of editing and subjective evaluations of perceived editing effort, the work examines what types of errors cause particular effort in post-editing, or even render the practice impracticable. As is likely to be the case in most human endeavours, also in post-editing it can be seen that there is a variation of effort between different people: some people work faster than others, and not all post-editors necessarily agree on how much or what kind of editing needs to be done. A further perspective of this dissertation, therefore, is an exploration of how the indicators of effort vary between different post-editors.

The dissertation consists of six previously published, peer-reviewed articles, and this introductory summary, which presents a general introduction to the articles, the theoretical and practical context of this work, the main results and discussion. Five of the articles forming this dissertation report original research, and involve analyses of post-editing data to examine questions related to post-editing effort as well as differences between post-editors. Three of the articles are co-authored, based on collaboration with other researchers in Finland and the United Kingdom. The sixth article is a survey presenting an overview of the research literature on post-editing, examining what we know so far about productivity, quality, and effort in post-editing.

The original articles are the following. They are referred to as Articles I–VI in the text, and are reproduced at the end of the print version of this dissertation. Rather than the chronological order of publication, this summary discusses the articles in the order that the research proceeded, which also follows the order of the research questions (see Section 1.2). Article I reports a literature survey that has been ongoing since the start of the dissertation project, although it was the last one to be published.

Article I	Maarit Koponen. Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort. In *Journal of Specialised Translation* 25, January 2016, pages 131–148.
Article II	Maarit Koponen. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190. Montréal, Canada, 2012.
Article III	Maarit Koponen, Wilker Aziz, Luciana Ramos and Lucia Specia. Post-editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*, pages 11–20. San Diego, California, 2012.
Article IV	Wilker Aziz, Maarit Koponen and Lucia Specia. Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia (eds). *Post-editing of Machine Translation: Processes and Application*, pages 170–199. Newcastle upon Tyne: Cambridge Scholars Publishing, 2014.
Article V	Maarit Koponen and Leena Salmi. On the correctness of machine translation: A machine translation post-editing task. In *Journal of Specialised Translation* 23, January 2015, pages 118–136.
Article VI	Maarit Koponen. This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 1-–9. Nice, France, 2013.

Research into the post-editing of machine translation is interdisciplinary by nature, combining topics, concepts, questions, methods and data from different fields. This dissertation also describes a mixed method study, with a combination of qualitative and quantitative approaches and multiple sets of data, approaching the question of post-editing effort from slightly different perspectives. On one hand, it relates to the field of language technology, particularly research on machine translation and automatic evaluation; on the other, it deals with translation studies, particularly the descriptive field of translation process research. The interdisciplinary nature can also be seen in that the published articles have different publication channels and target audiences: some are oriented towards a more technical audience, while others are more closely related to translation. This summary, then, aims to bring together the different perspectives, addressing both audiences.

The summary is structured as follows: Chapter 1 provides a general introduction to this dissertation project: the research problem, the objectives, and research questions. Chapter 2 presents the practical and theoretical context of the work as a whole, providing an overview of the current state of the practical use of post-editing and related research. Chapter 3 presents the research design, including the data and methodology of the research articles. Chapter 4 then describes the main results of the research articles with regard to the research questions presented. Finally, Chapter 5 discusses the theoretical and practical implications of these results, as well as an evaluation of the research design.

1.1 Background and research problem

The initial motivation for this dissertation arose from an interest in machine translation (MT), particularly from the perspective of a translator: Can MT be useful for a translator? When, and how? Around late 2011, this interest finally led me to a closer examination of post-editing of machine translation as a topic. While MT remained far from perfect except for some very limited contexts, increasing reports appeared of it being used to provide raw versions to be then corrected by translators. Around this time, a general surge of interest in researching post-editing was also apparent: the first Workshop on Post-editing Practice and Technology, bringing together researchers and practitioners in the field, was arranged in 2012 and they have continued annually.[1] Research published around that time did indeed suggest that post-editing sufficiently high-quality machine translation can increase productivity in terms of translation speed (for example, Plitt and Masselot 2010). As attested by the spreading practical use and the research reports, there was a purpose, then, where MT appeared "useful enough", to cite Koby (2001, 5).

Although the potential for MT to be useful in the post-editing scenario has been shown, its usefulness relies on the quality of the MT being high. If the raw machine translation is of poor quality, post-editing may end up taking even more time and effort than discarding the MT and translating from scratch. Therefore, one of the key questions in post-editing is estimating the amount of effort involved in post-editing and the amount of effort acceptable: When is using machine translation worth the effort? Accurate measurement of the actual effort is important as it addresses not only productivity, but also the working conditions of the post-editors, as argued by Thicke (2013, 9): the post-editors generally spend as much effort as needed to correct the MT, and are ultimately the ones paying the price for poor MT quality. Further, finding ways to decrease the effort involved in post-editing is important for both productivity and working conditions (see Krings 2001, 51).

The concept of *effort*, itself, is not straightforward. In his seminal work on PE effort, Hans P. Krings (2001) argues that the effort caused by post-editing in fact is comprised of three different aspects: perhaps most visible is the *time*

[1]For the most recent workshop in 2015 and links to the previous ones, see `http://wptp2015.postediting.org/`.

aspect of effort, or how long it takes for the post-editor to edit a given text or passage. The second aspect involves the *technical* effort needed to perform the corrections. But before corrections can be made, *cognitive* effort is required to identify the errors and plan the corrections. Various approaches have been taken to measure the effort involved in post-editing, commonly in the form of measuring post-editing time or the number of changes performed, in terms of words changed or keystrokes (for a more detailed description, see Section 2.3). These metrics capture the temporal and the technical aspects of effort, but the cognitive aspect is more difficult to capture.

The more detailed research questions of this dissertation started to take shape during an examination of the data analyzed for Article II. In addition to post-edited MT sentences, this dataset also contains effort-related evaluation scores reflecting the post-editors' assessments of the effort involved in post-editing each sentence (Callison-Burch et al. 2012). Although PE time and the number of changes have been shown overall to correspond relatively well with such effort-related evaluation scores, they do not always agree at the sentence level (Specia and Farzindar 2010; Sousa et al. 2011). Comparing the evaluation scores to the amount of editing performed in each sentence brought up cases where discrepancies could be observed. In some cases, the score indicated that post-editing would require a great deal of effort, but relatively little had in fact been changed – or vice versa. Similar observations can be made in datasets involving PE time and the number of changes: for some sentences, post-editing appears to take a longer time than would be predicted on the number of changes alone.

The discrepancies between the perception of PE effort or PE time and the amount of post-editing performed pointed to a need for a closer analysis of such cases. If the number of errors and edits were not sufficient to explain perceived effort, perhaps specific types of MT errors and edits could be the cause. The fact that not all errors are equal is well known, both in human translation (see, for example Bensoussan and Rosenhouse 1990), and in post-editing (Krings 2001, 179). In the context of post-editing, it can be argued that the features of sentences where PE effort is particularly high or particularly low would be of special interest for detailed investigation, as such features can help predict the usability of the MT (Krings 2001, 181–182). While some work has focused on investigation of the connections between effort, MT errors, and source text features (see Section 2.3.2), much remains to be done in this area. This dissertation contributes to this line of investigation through analyses of what features may be connected to increased PE effort.

A further point of interest arose in the planning stages of the study reported in Article III, originally suggested by my co-author Dr. Lucia Specia. Analyses of PE time and amount of change are naturally also affected by overall observations that post-editors differ in terms of how they approach the corrections during PE, the time they take to post-edit and the amount of editing they do, as well as the potential productivity increases (Plitt and Masselot 2010; Tatsumi and Roturier 2010; Sousa et al. 2011). This pointed to a need to investigate how much, and in what ways, post-editors differ from each other. This type of investigation would inform the analysis of how PE effort can even be generalized across different people. The differences between post-editors can be examined from two points of view: based on their processes during post-editing and based on their products, the final translations they produce. Although some work has been done in this area (see Section 2.4), for example in the investigation of how the post-editors' experience affects the effort indicators (Guerberof Arenas 2014b; Aranberri et al. 2014), work also remains to be done. This dissertation contributes to the research through analyses of multiple post-editors' process data as well as their final translations.

1.2 Objectives

The purpose of this dissertation is to explore the interaction between translators and machine translation in the post-editing context. The central questions investigated focus on the effort involved in post-editing. This dissertation approaches the question of effort not only based on commonly used indicators of PE effort, such as PE time and the number of changes made, but also the post-editors' perception of effort, as well as their choices related to what can be edited and how. The issue of PE effort, and particularly identifying situations where effort is high, is important for two reasons. Firstly, it may indicate particularly critical errors to be addressed in the development of MT systems. Secondly, it is necessary to accurately estimate the post-editors' workload involved in producing the final translation. In general, the goals of this dissertation can be divided into three separate yet interconnected issues.

The first issue is most closely focused on investigating MT error types potentially associated with low or high effort in PE. For the purposes of this dissertation, what constitutes an *error* follows the definition set out by Green (1982, 101): an error is "any feature of the translation which causes the posteditor to put pen to paper". Errors are thereby defined through the actions of the post-editor, and evaluating the quality of the post-editor's final product is not within the scope of this study.

The first goal of this dissertation is to identify situations that require particular effort from the post-editors, especially effort that may not be immediately evident in the number of changes performed. The question of particularly difficult edits is investigated in this work by analyzing errors found in the machine translations and their relationship with effort indicators like time and effort-related evaluations carried out by editors. A related second issue is explored in terms of the cognitive aspect of effort, namely the effort involved in recovering the source text meaning despite MT errors. More specifically, this dissertation aims to identify situations where the post-editors can detect and correct errors easily even without the source text, and situations where they cannot, or do not even attempt to, correct the MT. This goal is addressed by analyzing editors' choices and MT errors in a context where no source text is available.

The final issue explored in this dissertation relates to the PE processes employed by different post-editors and how they differ in terms of various effort indicators. This issue relates to the accurate assessment of effort, and particularly to certain assumptions about the PE process and the effort involved. To determine how to assess effort, we need to look into the post-editing processes of different people and the potential commonalities and differences found in their working processes and choices. This dissertation therefore aims to investigate the variation between editors by analyzing their process data – PE time, the way edits are made – as well as their subjective choices related to what is easiest to edit. This dissertation does not aim to make any suggestions or commentary on how the post-editors *should* approach the process of post-editing. Rather, the interest is on investigating the process data as-is.

These three main issues addressed in this dissertation are summarized in the following Research Questions and sub-questions:

RQ1 What types of errors and edits can be identified in sentences requiring much effort or little effort?
   (a) Using human evaluations as an indicator of effort?
   (b) Using post-editing time as an indicator of effort?

RQ2 How do errors affect cognitive effort when post-editing without a source text?

RQ3 How do different post-editors and their edits differ from each other?
   (a) In terms of effort indicators related to process data?
   (b) In terms of preferences related to editing choices?

Each Research Question is addressed in one or more of the peer-reviewed articles forming part of this dissertation. Table 1.1 shows how the questions and articles are connected. Article I addresses all three of the research questions by presenting a survey of the literature and research related to PEMT and particularly PE effort. RQ1 is addressed in Article II, Article III, and Article IV. Article II focuses on sub-question RQ1a, describing an experiment that involves human evaluations of effort. Article III and Article IV look into sub-question RQ1b, describing two experiments where PE time and other process data was used as an indicator of effort. RQ2 is addressed in Article V, which describes an experiment involving PE without source text. RQ3 is addressed in Article III and Article VI. Article III focuses on sub-question RQ3a and describes an analysis of process

Article	RQ1	RQ2	RQ3	Goal
I	X	X	X	Survey the research literature on PEMT and related research, with a focus on PE effort.
II	X			Investigate the relationship between edits and human evaluation of effort.
III	X		X	Investigate the relationship between edits and PE time. Investigate the differences between post-editors.
IV	X			Investigate the relationship between ST features and PE time.
V		X		Investigate the relationship between MT errors and effort in monolingual PE.
VI			X	Investigate the differences between post-editors and their edits.

Table 1.1: Research Questions addressed in each article

data from different post-editors. Article VI focuses on sub-question RQ3b and discusses an experiment involving post-editors' choices of what to edit and how.

Chapter 2
Theoretical background

This section describes the practical and scientific context of the dissertation. The chapter is organized as follows. Section 2.1 presents a brief overview of the history and current context of post-editing in practice. Section 2.2 presents an overview of research into the process of post-editing and its relationship to translation and revision processes. Section 2.3 provides a more detailed look into the question of post-editing effort and factors affecting it. Section 2.4 presents research related to differences observed between post-editors.

2.1 Background to post-editing in practice

Using computers to automatically translate texts from one language to another, or machine translation (MT), has been envisioned since at least the mid 1940s. The task of creating a fully automatic translation system turned out to be very complex, however. Even some seven decades later, although MT may occasionally achieve high quality on isolated sentences, no machine can yet produce a full text even approaching the quality of a human translator. Even if the goal of fully automatic high-quality MT appears unattainable, there are certain approaches that can make MT usable. They can be summarized as follows, partly following the categorization suggested by Krings (2001, 37). One approach is to reduce the potential ambiguity and problems in the source text either by restricting the source texts to be translated to specific text types and domains where the text features can be defined precisely (Krings' first strategy), or by pre-editing the source text according to specific simplification rules before MT (Krings' second strategy). Another approach would be to use interactive MT systems, where a human translator solves problems arising during machine translation (Krings' third strategy). These types of "human assisted MT" systems where the user resolves ambiguous cases but the system otherwise translates automatically do not appear to be in general use. The next approach is to use MT to provide a quick, rough idea of the general information content of the text in situations where high-quality text is not needed (Krings' fourth strategy). This use scenario is targeted by the commonly available free online MT systems (such as Google Translate). The final approach involves using the MT to produce a raw version to be edited by a human translator (Krings' fifth strategy). This last approach, generally referred to as *post-editing of machine translation* (PEMT) or *post-editing* (PE)[1], has attracted increasing interest over the past few years, and forms the

[1] There seems to be no agreement on whether the terms "post-editing" and "post-editor" should be written with or without the hyphen. Both forms have been in use since the introduction of the concept, and continue to be used by different authors. In this summary, as in my previous articles, I have chosen to use the hyphenated form.

central topic of this dissertation.

Although the surge of interest in MT and PE workflows appears recent, the idea is in fact not new. The difficulties related to fully automatic MT were recognized very early on, and according to Hutchins (1986, 31), the concept of a *post-editor* appears first in 1950, when it was used to describe a person who would select the correct translation from suggestions provided by a computer dictionary, and rearrange the word order in accordance with the target language rules. An early description of a PE workflow implemented at the RAND Corporation appears in Edmundson and Hayes (1958). However, early interest particularly in the US diminished after the report by the Automatic Language Processing Committee (ALPAC 1966) found post-editing not to be worth the effort in terms of time, quality and difficulty compared to human translation. For a more detailed historical overview of PE and related research from early suggestions in the 1950s up to the present, see García (2012). Development still continued to some extent, and MT systems combined with PE have been in use since the 1980s in large organizations like the European Union and the Pan-American Health Organization (García 2012, 295–299). Recent years have seen increasing interest in PE workflows in business contexts (see, for example Plitt and Masselot 2010; Zhechev 2014; Silva 2014).

It is difficult to determine precisely how widely PEMT is used at present. Some general trends are indicated by surveys carried out in the language industry. In January 2010, a survey carried out by the Translation Automation User Society reported that about half of the 75 language service providers surveyed from around the world were, at that point, offering PE services (TAUS 2010b, 13). A more recent report by TAUS (2014) does not give numbers on how many of the 80 companies surveyed are offering PEMT services. However, it states that various language service providers estimated that PEMT accounted for shares ranging from 5% up to 40% of their total translation production (van der Meer and Ruopp 2014, 37). As an example, van der Meer and Ruopp (2014) report that between July 2013 and July 2014, 50% of the translation technology provider Memsource's client projects were set up using MT as one translation source. Of the 438 stakeholders in the translation and localization field surveyed by Gaspari et al. (2015), 30% were using MT, and the majority (70%) of the MT users combined that with PE at least some of the time. Gaspari et al. (2015) also provide a summary of various translation industry surveys from the past few years, which indicate that the demand for PE services is growing, as is the number of language service providers implementing MT systems. PEMT is even predicted to become "the primary production process in translation" by 2020 (van der Meer and Ruopp 2014, 46).

The adoption rate of MT and PE workflows naturally varies in different countries and language pairs. In Finland, PE has not yet been extensively used, partly due to the small market. Partially the situation may also be connected to MT quality achievable with Finnish as one part of the language pair: as a morphologically rich language, Finnish has proven difficult for MT systems. A discussion of the particularities of the Finnish language and MT can be found in Koskenniemi et al. (2012, 59–61). The quality issues are reflected, for example, in the European Commission trials, where English-Finnish translators considered MT sufficient to suggest ideas for expressions, at best, or unusable and better replaced by translation from scratch (Leal Fontes 2013). This can be contrasted with more successful language pairs like French-Spanish, French-Italian and French-Portuguese, where most MT sentences were rated reusable (Leal Fontes 2013, 11).

In a recent survey focusing on the IT skills of Finnish translators, Mikhailov (2015) found that MT-related skills were not considered important by the translators themselves, and points out that this opinion is likely connected to the unavailability and low quality of MT systems involving Finnish. No explicit numbers are given on how many of the 238 respondents to the survey had used MT systems, but it is noted that most were only familiar with free online systems like Google Translate: only three respondents mentioned having used some other system (Mikhailov 2015, 111, endnote 5). Development work is, however, being carried out by Finnish language service providers (for example, Nieminen 2013), and at least some freelance translators are adopting PE workflows (Rosti and Jaatinen 2015).

The driving force behind this interest in implementing PE workflows most likely relates to the hope of increasing productivity: translating more pages in a shorter time, with lower costs. Recent studies have indeed shown that MT in business contexts can increase the productivity of translators at least for some language pairs. For example, an estimate is offered by Robert (2013, 32), who states PEMT can increase the number of words translated by a professional post-editor by 2,000 to 3,500 words per day. In terms of relative numbers, Plitt and Masselot (2010, 10) found an average increase of 74% in the number of words translated per hour. As Guerberof Arenas (2010, 3) points out, however, the actual numbers vary in different projects and language pairs (see also Plitt and Masselot 2010; Zhechev 2014). It is also important to note, as García (2011, 228) does, that these promises of large increases in productivity also depend on circumstances like professional, experienced post-editors using domain-specific MT systems, and source texts that may have been pre-edited for MT. Productivity increases do not necessarily appear, at least not to such an extent, in situations where these conditions do not hold. No significant increases were found by Carl et al. (2011) and García (2010), whose studies involved PE by non-professional participants.

2.2 The post-editing process

To investigate PEMT, it is necessary to define the practice, and the similarities and differences compared to translation practices not involving the use of MT. Koby (2001, 1) states that post-editing is "usually understood as a human being (normally a translator) comparing a source text with the machine translation and making changes to it to make it acceptable for its intended purpose". This appears to reflect the general understanding of the term, and the process behind it, but some exceptions may occur. For instance, the possibility of post-editing MT monolingually, without the source text – which was the assumed condition of the early researchers like Edmundson and Hayes (1958) – has been explored (for example, Koehn 2010; Hu et al. 2011; Mitchell et al. 2013). This monolingual scenario is discussed further in sub-section 2.2.2.

As another departure from the general definition provided by Koby, the post-editor is not always a translator, at least not a professional one. On the research side, many experiments have involved semi-professional or non-professional translators as post-editors (for example, García 2010; Carl et al. 2011). Practical scenarios where non-professionals act as post-editors may involve the translation of content that is unlikely to be translated otherwise, but might be useful for information purposes. Suitable texts for such non-professional crowd-sourced PE may be, for example, user-generated context like technical forum or social media posts (see Mitchell et al. 2014; O'Curran 2014). As noted by O'Curran (2014, 53), the language skills of the post-editors still have to be vetted carefully. For this reason, the term *post-editor* is used in this summary as a general term to mean anyone who carries out the task of post-editing MT, without reference to their potential professional status.

A noteworthy point in Koby's definition concerns the purpose of PEMT: to make the translation "acceptable for the intended purpose". This refers to the idea that the goal is not necessarily always to produce a perfectly polished text. For instance, in the case of user-generated content, a lower quality translation is more useful than no translation at all for a person searching for information. In certain other situations, however, such information quality would not be sufficient. Instead, quality similar to professional human translation would be required. As an example, O'Curran (2014, 53) specifies that any "high-end, high-visibility content" and content "expected to convey the company image" is not suitable for PEMT, and human translation is to be preferred. The issue of quality expectations is raised also by Koby (2001, 8), who notes that human translators generally work with the standard of having to create a perfect text, but the actual end users may have a lower expectation. While the "good enough" standard may appear unintuitive for the translator, Koby (2001, 8) argues that the user's need is the ultimate factor determining the quality level.

The characteristics of two such quality levels are defined in a set of PE guidelines drafted by the think-tank TAUS (formerly Translation Automation User Society) with the intention of helping customers and language service providers set clear expectations and instruct post-editors. The first level, termed "good enough" quality, is defined as comprehensible and accurate so that it conveys the meaning of the source text, but not necessarily 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 the text does not contain any 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 good. (TAUS 2010a, 3–4)

Another way of distinguishing between different types of PE goals is to define the level of intervention required. According to Krings (2001, 45), such a distinction was introduced first by Löffler-Laurian (1986), who describes two PE levels: *post-édition rapide* and *post-édition conventionelle* in the original French. The English edition of Krings (2001) uses the terms *partial* and *complete* post-editing for the distinction, while Allen (2003) uses a slightly different definition of three PE levels (rapid, minimal and full). O'Curran (2014, 52) defines three levels, where *light* PE is simply a "sanity check" to ensure there are no misrepresentations or offensive content, *medium* PE emphasizes meaning and readability but fluency and style are not expected to be perfect, and finally *full* PE ensures correct grammar, fluency, terminology, style and voice.

The specifications for the expected quality of the post-edited text, on one hand, and the level of PE intervention, on the other, are naturally linked to some extent. However, the relationship is not necessarily straightforward. As pointed out in TAUS (2010a), the raw MT quality also plays a role: if the MT is of very high quality, light PE might be sufficient to bring it to publishable quality, but if it the MT is very poor, even full PE may not help to make it even good enough.

The above definition offered by Koby (2001) also provides an overview of the overall tasks of the post-editor: to compare the MT to the ST and to make the necessary corrections to produce a target text (TT). It is, however, necessary to describe the actual process of PE in more detail. This need for a general process description is brought up by Krings (2001, 164), who further notes that the general process can then be divided into subprocesses, and examined in terms of individual micro-level processes as well as the way they are combined into macro-level processes at a given phase of the overall process. To some extent, the processes of PEMT are likely to be shared with human translation (where MT is not used): there are processes related to reading the source and target (and MT in the case of PE), comparison and evaluation of the ST, TT and MT, and decision-making.

On a very general level, the processes involved in PEMT can be viewed in terms of Englund Dimitrova's (2005) model of translation as text production, which is an adaptation of earlier work describing the writing process in general. According to Englund Dimitrova (2005, 19–20), this model consists of three main cognitive processes: planning, text generation, and revision. The processes are at each moment during writing influenced by the task environment, which includes the (writing or translation) assignment, topic, audience, and the text produced so far. They are also influenced by the long term memory of the writer, which contains knowledge of the topic and audience, and writing plans. In the case of translation, the ST becomes part of the task environment, and source language (SL) and target language (TL) knowledge is also contained in the long term memory (Englund Dimitrova 2005, 19–20). Although three main cognitive processes are described in the model and many researchers have identified very similar phases, Englund Dimitrova (2005, 22) does point out that there is some difficulty involved in dividing the process into exact stages, because all the stages involve many cognitive processes: reading the ST or TT, writing, transfer from SL to TL and evaluating what has been written so far.

In order to extend this model of translation and text production to PEMT, the MT has to be added to the task environment together with the ST and the (edited) TT produced so far by the post-editor. Long term memory would also include, at least in the more general bilingual PE scenario, SL and TL knowledge, and perhaps some specific knowledge of how PE is carried out. The kinds of PE guidelines and quality level specifications discussed might form part of the writing assignment. The main cognitive processes of planning, text generation, and revision, as well as the processes they involve – reading, writing, transfer, and evaluation – apply to PEMT, as well. However, the addition of the MT to the situation, and the consequent changes related to the task environment and long term memory, are likely to affect the process in some manner. The differences between human translation and PEMT processes are discussed in more detail in the following subsection.

2.2.1 Post-editing compared to translation and revision

As both the task environment and the working memory are seen to influence the processes, and these factors are affected by the addition of the MT, PEMT processes are also likely to differ compared to translation. Krings (2001, 165–166) suggests two ways that the nature of MT texts may influence the process: firstly, the defects present in MT may break up normal reading patterns, and secondly, the equivalency search processes may be affected by suggestions already provided by the MT. Overall, the results of Krings's comparison of translation and PEMT suggest that although there were no real differences in the *types* of processes or subprocesses, the *distribution* of processes was different (2001, 545).

Translation, like text production in general, is processed as segments of a limited span, generally shorter than a sentence. This segmentation is due to limitations of working memory, which allows only a limited amount of information to be held in the working memory at any given moment (Englund Dimitrova 2005, 27). In Translation Studies, these segments are generally termed translation units, and according to Alves et al. (2010, 121–122), they are cognitive units that can be observed as continuous "chunks of activity" happening between pauses in the writing process. More specifically, the translation units are not necessarily grammatical units. Rather, they reflect a certain segment of the ST that the translator's attention focuses on, which can be captured by mapping the TT segments produced in continuous writing between pauses onto the corresponding ST segments (Alves et al. 2010, 124–125).

Adding the MT to the situation may then change this segmentation in the reading and processing of the ST. As Krings (2001, 166–167) observes, the post-editor has two possible ways of reading the texts in the PEMT situation: either mainly to orient to the MT and only check the ST when the MT is unclear or incorrect, or mainly to orient to the ST and then use the MT as parts to be used in their own translation. Further, Krings (2001, 168) argues that the often defective nature of MT may direct these processes to a lower level, so that MT is read and analysed at a lower linguistic level, and the need to compare the ST and MT (and TT) may further limit the amount of text that can be processed as a unit. The results reported by Krings (2001, 534) show that PEMT in fact involved a larger proportion of ST related processes than translation. Combined with the larger number of attention shifts identified, this suggests that needing to deal with splitting the attention between three texts indeed limits the length of the segment that can be held in working memory, making it necessary to read the ST more often (Krings 2001). A more recent study by Carl et al. (2011, 140) found that visual attention was much more focused on the TT area during PEMT than during translation, while during translation, fixations in the ST were more frequent and longer. Along similar lines as Krings, Carl et al. (2011, 140) discuss how the eye movements of the post-editors reflect a process where they first read the MT, then the ST and then the MT again, while the patterns for translation could mean that the ST is understood more deeply.

In applying the text production model to translation, Englund Dimitrova (2005, 25–26) argues that one of the key differences is that, in translation, the contents and many aspects of structure are already set and limited by the source text. This naturally affects the planning stage: planning has already been done by the ST writer. In the case of PEMT, the addition of the MT to the task environment then introduces a new factor that may again direct the post-editor's planning and choices. The directing effect can be seen, for example, in that fewer variant forms are produced and considered during PEMT than during translation Krings (2001, 537). This may in some cases ease the planning stage, but there are also some dangers to the seductiveness of these suggestions: MT tends to be overly literal and project the ST structures onto the TT, making it awkward, and a post-editor may become so accustomed to seeing this type of text that they no longer see the errors (Koby 2001, 10–11). Krings (2001, 537) also states that PEMT triggers less processes related to TT evaluation, suggesting that the task may make the post-editors less critical of the alternatives. This is supported by the more recent pilot study by Čulo et al. (2014) examining certain specific cases of lexical or structural interference in PEMT. Interference was found more commonly in post-edited texts than those translated from scratch, and based on these observations, they suggest the reason is twofold: firstly, dealing with the often defective MT may lead to unidiomatic choices going unnoticed, and secondly the task description may lead the post-editor to decide that the unidiomatic choice is understandable and therefore good enough for the purpose (Čulo et al. 2014, 213–214).

Various studies have aimed to investigate the question of whether the directing effect of the MT leads to a lower quality of the end product. The results so far suggest that this is not necessarily the case. Fiederer and O'Brien (2009, 62–63) approached the question by comparing post-edited and manually translated versions, and found that the PE versions were rated as similar to manual translation in terms of the clarity of the sentences, worse in terms of style (appropriateness and idiomaticity), but in fact slightly better in terms of the accuracy of meaning. Also in the study by Carl et al. (2011), post-edited versions were ranked slightly better. Studies utilizing error-based quality evaluation have shown similar findings. When assessed by Autodesk's quality assurance team according to the criteria applied to all their translations, both manual translations and post-edited texts were all rated as acceptable for publication, but the manually translated texts in fact contained more sentences that were flagged as needing corrections (Plitt and Masselot 2010, 14–15). Using the guidelines created by the Australian National Accreditation Authority for Translators and Interpreters, García (2010) similarly found that the PE versions received slightly higher average marks.

One of the key quality issues discussed in context of post-edited texts relates to the "literalness" of MT. The term literal translation is generally used to describe translations that are maximally close to the ST meaning and structures while still conforming to TL grammar, and they are generally viewed negatively in Translation Studies (Englund Dimitrova 2005, 51–52). In the studies cited above, it was seen that in some cases evaluators even preferred the post-edited versions. An interesting suggestion for why this might be comes from Koby (2001). He notes that "nonlinguist" users cannot always tell the difference between MT and human translations, and sometimes even deem the human translation to be the one that is incorrect or not fluent enough. As an explanation, Koby (2001, 10) offers that if a MT is good enough to get the terminology and individual phrases correct, it may even appear more consistent than a human translation. It may also be that the nonlinguist readers are not as critical of literal translations as translators themselves. In fact, Englund Dimitrova (2005, 52) does bring up the fact that also some Translation Studies scholars argue for the acceptability of literal translations.

In addition to translation, we can compare PEMT to *revising*, which refers to the "process of checking a draft translation for errors and making appropriate amendments" (Mossop 2007, 202). Mossop (2007, 125) categorizes the potential errors checked for into four groups: meaning transfer, content, language, and presentation (for example, layout). On a general level, these features are similar to what the post-editor also has to check – for example, the accuracy and completeness of meaning or idiomatic and grammatically correct language. There are, however, also differences. According to Koby (2001, 6–7), one of the key differences between PEMT and revision is that the human translator and the reviser both draw from a shared background of extratextual knowledge of the source and target cultures, texts, and languages, as well as differences between them – knowledge which the MT system does not have. The reviser's task mainly involves checking for inadvertent omissions or misunderstandings, and a misunderstanding by a human translator would likely affect the text from that point onward (unless the translator comes across a passage that cannot be reconciled with the misunderstanding and goes back to check). The MT system, on the other hand, would make more localized errors related to mistranslated words and structures. (Koby 2001, 7)

The processes of PEMT and revision also share some similarities, since both involve comparing a TT to the ST and making corrections. Both offer the reviser or post-editor two choices when faced with a problematic passage: to either rewrite the passage completely, or to revise it (Englund Dimitrova 2005, 31). Englund Dimitrova (2005, 31) argues that revision is, in fact, the more difficult option, because it is necessary to first diagnose the problem and then find a way to correct it. In this sense, PEMT may in fact be easier than revision. When faced with the choice of whether to rewrite completely or to revise, practical experiences have shown that post-editors of MT have less inhibitions than revisers of HT, in that they make larger changes more rapidly and effectively (Koby 2001, 10). Krings (2001, 61–62) brings up a similar point that there is a relative psychological ease related to PEMT in that there is no concern about insulting the machine.

The use of MT as an information source in translation is also related to the common use of translation memory systems. A translation memory (TM) is a database containing previously done translations saved as source and target segments (generally sentences), which are then compared to the new source text and offered to the translator to be re-used if the source segments match fully or partially (Paulsen Christensen and Schjoldager 2010, 3). If the ST segments are not fully identical, the translator then needs to check the differences and perform the necessary changes. Partial TM matches are given a percentage value to indicate their similarity to the new ST segment, based on a comparison of the strings of characters. For a more detailed overview of TM technology and related research, see Paulsen Christensen and Schjoldager (2010). In practice, it is becoming more common to combine suggestions offered by TM and MT (Guerberof Arenas 2014a, 165).

As Teixeira (2014a, 120) points out, the use of TM and MT are similar in the sense that both TM and MT provide suggestions that can help the translator or post-editor to generate viable solutions. Both also share the similarity in that, similarly to revision, they involve comparing the TT to ST and making the necessary corrections. In terms of productivity and quality, high-quality MT suggestions have been found to be very similar to TM matches with match values of 80–90% (O'Brien 2006a, 199–200) or 85–94% (Guerberof Arenas 2014a, 184). Guerberof Arenas (2014a) also reports that both TM and MT improved productivity over translation from scratch, and sentences translated using TM or MT contained fewer errors.

Despite some similarities between TM and MT, there are also differences. For one, the things that likely need to be corrected are different. In partial TM matches, the ST segments differ in some way, and the TT segment needs to be adjusted accordingly. TM matches, however, are unlikely to contain the kinds of lexical and grammatical errors found in MT. For TM matches, the software generally also offers information about the origin as well as the quality of the TM suggestion, in the form of the match percent. For MT, there is generally no indication of the origin (beyond potentially the system name) and how good the translation is likely to be (Teixeira 2011, 108). Often the TM system also highlights the textual differences in a TM match, providing a visual clue for the translator what needs to be changed (Teixeira 2014a, 118). This type of information is not available for the MT. Teixeira (2014b, 46) also points out that the normal working modes differ: TM is generally used interactively, with suggestions being shown segment-by-segment, while with MT, the entire text is generally pretranslated before starting the PE. However, the pretranslation mode is also sometimes used with TM (Paulsen Christensen and Schjoldager 2010, 3). An interesting difference between these two workflows is brought up by Teixeira (2014b), based on an investigation of both TM and MT segments used in both modes. In the interactive mode, the participants did more iterations of correcting and then revising their own corrections, while in the pre-translated condition they tended to make only one round of corrections, regardless of whether the suggestions were from TM or MT (Teixeira 2014b, 51–52).

This sub-section has examined the ways that PEMT is similar to, and different from, translation and revision. One of the central features of these three tasks is that they all involve some source text that is used as a point of departure for producing the final translation. The next sub-section will turn to a special scenario of PEMT: monolingual post-editing, where the source text is not available. This task differs considerably from translation and also revision, although revision can sometimes involve "unilingual re-reading" (Mossop 2007, 110), where the ST is not used.

2.2.2 Post-editing without source text

In practice, PEMT is mainly carried out in scenarios where the post-editor produces a TT based on the ST and MT. However, studies have also explored another possibility, sometimes termed "monolingual PE", where the post-editor works based on the MT alone, without ST. This approach was suggested already by the earliest writers like Edmundson and Hayes (1958), who assumed that the post-editor would utilize the machine-translated text and a grammar code indicating the part of speech, case, number and other details of every word, but would have no need to even see the ST. Although this attempt may appear "counterintuitive", as Koby (2001, 2) puts it, there are some reasons why a monolingual process might be useful. Koby (2001, 1) suggests the lower hiring costs of monolingual post-editors as one reason. Other reasons examined in more recent studies involve, for example, translation in situations where professional translation is not likely to be used, such as translation of user-generated content like forum posts (Mitchell et al. 2013).

In this monolingual scenario, the central question is whether, and to what extent, it is even possible for a post-editor to interpret the meaning and correct errors without the ST. This question was first taken up by Krings (2001), who observed some improvement in that compared to raw MT, PE without ST improved sentence-level ratings given by translators by nearly one point. However, the scores remained far from perfect, and there was great variation between improvements in different sentences. (Krings 2001, 274) In addition to the evaluation of overall improvement, Krings (2001, 182–183) argues that it is necessary to investigate which types of errors the post-editors are able to compensate for and which are impossible. Some errors, like word order, incorrect punctuation and errors defined as relating to stylistic or coherence issues were corrected over 90% of the time, whereas errors related to incorrect parts-of-speech were successfully corrected in only 54% of the cases (Krings 2001, 272). Krings (2001, 275) also points out that although overall, 79% of errors were corrected, the remaining errors seriously altered the meaning of the sentences. Interestingly, (Krings 2001, 532) found that PE without ST was slightly faster than bilingual PE or human translation, but did not consider it a viable practice in light of the observations related to success.

More recently, the possibility of PE without ST was again addressed by Koehn (2010). Instead of a rating scale or error-based evaluation, the evaluation was performed on a sentence-level "correct/incorrect" standard, where a correct sentence was defined as "a fluent translation that contains the same meaning in the document context" (Koehn 2010, 541). The approach was intended also as a way for evaluating MT quality: whether the quality was sufficient to convey the source meaning. A similar approach was also utilized in the MT evaluation campaign of the Fifth Workshop on Statistical Machine Translation (Callison-Burch et al. 2010). Both of these studies showed great variation in the percentages of successfully corrected sentences, depending on the language pair and MT system used. Results reported by Koehn (2010) ranged from 26% to 35% of sentences; in Callison-Burch et al. (2010, 28) the percentage of correct sentences ranged from less than 10% in the worst case to as high as 80% in the best.

This definition of correctness provides little information about whether the incorrect sentences were defective with regard to language or meaning. Other research has, however, attempted to investigate this difference with evaluations that separate fluency of language and adequacy of meaning. One such example is Hu et al. (2011), who studied PE without ST focusing on text messages for emergency responders. In terms of meaning, 24% to 39% of sentences (depending on system and test set) achieved the highest adequacy score as rated by two evaluators, while the highest fluency score ranged from under 10% to over 30% (Hu et al. 2011, 401–403). From these results, it appears that more language errors remained than meaning errors. The situation appears slightly more complicated in the results reported by Mitchell et al. (2013), who used three separate evaluation scales for fluency, comprehensibility of meaning (regardless of whether the meaning is correct), and fidelity of meaning. In one language pair evaluated, fluency scores improved in 67% of sentences and comprehensibility scores in 57%. In the other language pair, the corresponding improvements were 63% and 49%. The fidelity scores, which reflect the correctness of meaning, improved in 43% of sentences in one language pair but in 67% of sentences in the other. In some cases, the score remained the same, and in some, the PE score was even lower than the score for raw MT. (Mitchell et al. 2013, 37–39) It should be noted that Mitchell et al. (2013) do not report the precise scores achieved by the MT or PE sentences, but rather whether PE improved the score or not. It is therefore not possible to comment on the number of language and meaning errors in their results. Focusing specifically on adequacy of meaning, Schwartz et al. (2014, 190) report that 63% of the sentences post-edited without ST were judged to be fully correct.

A further significant question related to the success of PE without ST is, of course, how it compares to PE with ST, or to human translation. Koehn (2010, 544) reports that translators working with the ST achieved much higher percentages for correct sentences (61% to 66%) compared to PE without ST when evaluated against the same criteria. Mitchell et al. (2013) also compared PE with and without ST. In their evaluation, both achieved very similar percentages of improved fluency scores (63% for both in one language pair, 67% without ST and 70% with ST in the other). Comprehensibility scores improved more often when the ST was available (57% vs 64% in one language pair, 49% vs 63% in the other). The fidelity scores are an interesting case. PE without ST was found to improve the fidelity scores less than PE with ST (43% vs 56% of sentences) in one language pair, but in the other language pair, PE without ST surprisingly improved the fidelity score slightly more often (67% of sentences) than PE with ST (64%). (Mitchell et al. 2013, 37–39) As noted, however, these percentages represent the number of cases where the score improved compared to the MT, and do not necessarily tell us in which case the actual scores were higher. In fact, examining the actual scores more closely, Mitchell et al. (2013, 40) note that average fidelity scores were higher for PE with ST, and the improvement achieved by PE without ST compared to MT was sometimes much smaller.

It is interesting to note that even the sentences produced with the help of the ST do not achieve perfect scores. In the study by Mitchell et al. (2013), the post-editors were non-professional volunteers with varying language skills, which may have affected their results. On the other hand, in Koehn's (2010) study the human translations used for comparison had been been produced by professional translators. Koehn (2010, 542–544) reports that although some mistakes were found in the human translated sentences, in other cases it was not clear why the evaluators had marked specific sentences as incorrect.

In addition to considerable variation in the success of PE without ST depending on the language pair and MT system used, studies also show great variation between different post-editors (see Koehn 2010; Mitchell et al. 2013; Schwartz et al. 2014; Schwartz 2014). A commonly suggested reason for these differences is related to the post-editors' knowledge of the subject area. This point has also been made by Krings (2001, 170), who states that extra-lingual knowledge becomes the most important source of knowledge if there is no well-formed source text. Therefore, post-editors with expertise in the subject area are likely to succeed better in PE without ST. Expert knowledge and PE without ST was investigated by Schwartz (2014) in a case study where a machine translated scientific text was post-edited by an expert in the scientific domain in question without knowledge of the ST. Nearly all (95.9%) of the sentences were judged to fully convey the meaning of the source text, and all but one of the remaining sentences were rated as mostly correct (Schwartz 2014, 38). Although the results are based on a case study involving only one post-editor and one text, it does suggest good potential when expert knowledge is available.

2.3 Post-editing effort

A key question in evaluating the viability of PEMT as a practice relates to how much effort it involves, particularly compared to human translation. How PE effort can be measured, how it compares to translation effort, and how effort can be reduced, are therefore central questions for PE research. Much of the interest, particularly on the industry side, focuses on the potential for saving time (and thereby money) through the use of PEMT. The PE time, meaning time spent in correcting machine translations, can be seen as “the most visible and economically most important aspect of post-editing effort” (Krings 2001, 178), but it reflects only one part of the effort involved. Post-editing time is formed by the technical effort involved in keyboard strokes and mouse clicks used to perform the corrections needed, as well as the cognitive effort involved in first detecting the errors and planning the corrections. These three separate but interlinked dimensions form the *temporal*, *technical*, and *cognitive* aspects of effort, as defined by Krings (2001, 178).

It should be noted that there is some terminological variation with regard to cognitive effort: in the literature related to PEMT, the terms *cognitive effort* and *cognitive load* both appear. Vieira (2014, 189) discusses the background of these terms in relation to educational psychology, pointing out that in that field, cognitive load is generally considered to be a wider concept, of which cognitive (or mental) effort forms one part. Similarly as in Vieira (2014), this summary uses the term *cognitive effort* in the sense defined by Krings (2001, 179) to refer to the “extent of cognitive processes”.

Of these three dimensions of effort, Krings (2001, 179) sees the cognitive aspect as the decisive factor determining PE effort. This can be connected to the argument that the use of MT carries a “penalty” to begin with since it is first necessary to determine whether the MT is usable or not. The usability may not always be obvious, but rather may take several comparisons between the ST and MT. This can be seen, for example, in Green’s discussion of “grey areas” which cause most trouble for the post-editors. The grey areas are described as cases where a machine translated sentence is in fact reasonably good, but contains “doubtful translations and near misses” (Green 1982, 102), meaning that the post-editor must make subjective decisions about whether corrections are needed, and if so, how extensive they should be. A similar issue is raised by Krings (2001, 539), who also found that medium-quality sentences in fact involved more effort than poor ones, and similarly attributes the result to constantly needing to compare the ST and MT. As an additional explanation, Englund Dimitrova (2005, 32) suggests that focused repair procedures, which these medium-quality sentences are more likely to require, are considered to be cognitively more demanding than discarding a poor translation and rewriting.

The details of what, precisely, constitutes effort have received much interest recently. The next sections take a closer look into how various researchers have approached PE effort and ways to measure it, as well as findings related to factors potentially affecting PE effort.

2.3.1 Approaches to measuring post-editing effort

As the temporal aspect of PE effort is important for practical scenarios, PE time has been a commonly used indicator of effort. In practice, it is generally the determining factor as to whether PEMT is feasible, and PE time is often examined relative to translation time, comparing whether the use of PEMT leads to a higher number of words translated in a given timeframe (for example, Plitt and Masselot 2010) However, collecting accurate time information is not necessarily easy in practical work contexts: specialized tools for collecting time data may not be available, and self-reports by translators or post-editors may not be detailed enough (Moran et al. 2014).

Krings (2001, 180) also argues that the measurement of "absolute PE effort" in terms of the time taken to edit a given text, passage or sentence, is useful for comparing different MT versions and also for comparing different sentences in order to form conclusions about the types of deficiencies in the sentences. Such comparisons have been carried out by various researchers attempting to define PE effort. Temnikova and Orasan (2009) contrasted the time translators spent fixing translations for texts produced according to a controlled language, versus translations produced using non-controlled language (see also Temnikova 2010). Specia (2011) investigated whether it is possible to automatically predict PE effort by comparing sentences predicted to be good and average quality sentences, and showed that MT sentences predicted to be good were post-edited much faster.

One of the early approaches to measuring PE effort, and particularly the cognitive aspect of effort, involved the use of Think Aloud (TA), where the post-editors verbally report their actions during the PE process. Krings (2001, 532–533) used metrics related to the TA verbalizations, such as the frequency of attention shifts between ST, MT, TT, and other resources, to examine cognitive effort with the assumption that more frequent shifts indicate a more fragmented process and therefore increased cognitive load due to more frequent comparison processes. While the TA methodology was widely used in translation process research in the 1990s and early 2000s, certain weaknesses have been identified. Thinking aloud can only capture the conscious part of cognitive processing, and has the effect of slowing down the process and potentially even changing the cognitive processing involved (for a more detailed discussion, see O'Brien 2005).

PE effort has also been investigated based on subjective effort assessments provided by post-editors. For this purpose, various evaluation scales specifically focused on perceived PE effort have been suggested. Specia et al. (2010a, 3376) describe a four-point scale for estimating the need for PE, where 1 indicates that a sentence requires complete retranslation, while 4 indicates that the sentence is “fit for purpose” without any editing. The same scale is used in Sousa et al. (2011) and Specia (2011). A different five-point scale was used in the evaluation task of the 2012 Seventh Workshop on Statistical Machine Translation, where evaluators were asked to evaluate how much of the sentence needed to be edited (Callison-Burch et al. 2012). Lacruz et al. (2014, 79) define another five-point scale which addresses the usability of individual sentences, from 1 indicating totally incomprehensible “gibberish” to 5 indicating a correct or nearly correct sentence.

Manual scores are often considered complicated because they are not repeatable and different evaluators do not necessarily agree in their assessments. Sousa et al. (2011) has, however, shown a connection between them and PE time: sentences requiring less time to edit are more often tagged as requiring low effort by evaluators. Human evaluations are also considered the most important metric in large-scale MT evaluations (for example, Callison-Burch et al. 2012). Such evaluations are, however, labour-intensive and therefore not feasible for situations where fast and frequent assessments are needed.

As a proxy for the more time-consuming human evaluations, certain automatic metrics have been developed. While the specifics of the metrics vary, in general terms they compare the word-level changes between the raw MT version and the post-edited version of a sentence based on string matching. One of the most commonly used metrics specifically for PE evaluation is the “Human-targeted Translation Edit Rate” or HTER (Snover et al. 2006). This metric compares the MT and PE versions of a sentence and computes the minimum number of word-level changes between them. These changes, termed *edit operations*, can be deletions, insertions or substitutions of words, or changes to word order. The number of edit operations is then divided by the number of words in the post-edited sentence, giving a sentence-level score called the *edit distance*. A HTER score of 0 indicates no changes have been performed, and a score of 1 indicates the entire sentence has been rewritten. For some languages, extensions have been made for recognizing word stems, synonyms and paraphrases (Snover et al. 2010). Combining interrelated edit operations (for example, editing an adjective attribute’s gender due to changing its head noun) instead of treating each edited word separately has been suggested as an alternative by Blain et al. (2011).

The benefit of HTER and other similar metrics is that they allow rapid, automated evaluations of the number and rough type of edits regardless of the languages involved. On the other hand, while the correlation between these edit distance metrics and time or human evaluations has been claimed to be good, they do not always accurately reflect these indicators of effort. In an investigation of the correlation between PE time and some edit distance metrics, Tatsumi (2009) found that these two measures do not always correspond to each other, and offered some variables such as source sentence length and structure as well as specific types of errors as possible explanations.

Another possible edit-based method is Choice Network Analysis (CNA). This approach was suggested originally by Campbell (2000) as a way to compare different translation versions created by multiple translators, following the assumption that source text items with multiple different translations indicate cognitive difficulty. In the context of PEMT, this method was investigated by O'Brien (2005), who compared several post-editors' versions to identify cases where their edits created multiple different variants. The number of variants available to, and considered by, the translator or post-editor are also mentioned as indicators of cognitive effort by Krings (2001, 536–537) and Englund Dimitrova (2005, 26). In a slightly different but related approach, variation was also investigated by Tatsumi et al. (2012) who investigated how often different post-editors created multiple successive PE versions of the same source sentence in crowd-sourced PE, and found that most sentences had only one PE version.

The way the PE task proceeds in terms of writing offers another possibility for examining potential effort. The linearity of writing, meaning how long the writing process proceeds without interruption is taken as an indicator of effort by Krings (2001, 538). Furthermore, examining the writing process as keystrokes recorded through keylogging provides information about the technical effort. Measuring the keystrokes used by the post-editor tells us how much typing effort was needed to perform the corrections (see O'Brien 2005; Carl et al. 2011; Elming et al. 2014; Lacruz et al. 2014). However, Englund Dimitrova (2005, 75) states that although keylogging makes it possible to follow patterns like pauses, segmentation, and revisions, and therefore make some inferences about the cognitive processes, the actual typing is not the "dominant occupation" during translation or PE: most of the time goes to other tasks like reading. Nevertheless, this type of data makes it possible to track to some extent when during a process a decision is made and whether there appear to be any problems associated with it. Although, as Krings (2001, 530) also points out, there is no necessary connection between the amount of editing (or number of revisions) the post-editor made, and the number of potential alternatives considered, which means the cognitive effort is not necessarily obvious.

Pauses found in the keylogging data recorded during PE have attracted attention as a way of identifying potential problems, and thereby cognitive effort. O'Brien (2005) discusses an analysis of sample data which shows a correlation between certain source text features identified as potentially difficult and pauses, suggesting increased cognitive effort. On the other hand, a subsequent study by O'Brien (2006b) showed there were difficulties in connecting the location and duration of pauses to the cognitive processing. While O'Brien (2006b) and much of the earlier research involves identifying long pauses, a different approach has been suggested by Lacruz et al. (2012), who use clusters of short pauses in post-edited sentences (see also Lacruz and Shreve 2014; Lacruz et al. 2014). The hypothesis behind this approach is that sentences requiring more cognitive PE effort would contain a higher density of short pauses than sentences requiring little effort (Lacruz and Shreve 2014, 263). Interpreting pauses is difficult, however, since it is not known precisely what is happening during the period of inactivity on the keyboard. For example, Englund Dimitrova (2005, 27–28) points out that while planning and problem-solving during translation lead to pauses, a pause can also indicate that the translator is reading and evaluating what has already been written (see also O'Brien 2006b). Alves et al. (2010, 129) list planning, consulting external resources for alternatives, assessing the text already produced, or reading new text as possible reasons for pauses.

The development of eyetracking technology has also enabled the use of gaze data to measure effort. The eye tracker uses infra-red light reflected from the eye of a person looking at a screen to determine where the person is looking, and the reading process can be examined by tracking the reader's eye movements on the screen (Rayner et al. 2012). The general assumption behind is that the eye movements reflect the processes of the mind during reading, and that the spot where the reader's gaze fixates indicates the focus of attention (Hyönä 1993, 10–11). Furthermore, long duration of fixations or the occurrence of multiple fixations in the same spot (for example, a word) are assumed to indicate increased cognitive effort (Carl et al. 2011, 138). Commonly used measures are the fixation count, which reflects the number of times that the reader's gaze fixates on the same word or segment, and average or total fixation duration, which reflect how long the combined fixations were (see Carl et al. 2011; Vieira 2014).

In studies involving reading MT, eye tracking has been used to investigate the relationship between effort and MT quality by comparing the effort measured by gaze data with human evaluations of the MT quality (Doherty and O'Brien 2009; Doherty et al. 2010) or with automatic MT quality measures (O'Brien 2011). In PEMT studies, eye tracking has been used to compare the differences in fixations during post-editing and manual translation, where Carl et al. (2011, 140) found that the number and duration of fixations indicated more effort on the TT side during PEMT. Specifically focusing on cognitive effort, fixation counts and durations have been used by Vieira (2014) to identify potential ST features leading to increased effort.

The growing interest in collecting data for evaluation of PE effort has led to the development of various tools (for example, Aziz et al. 2012; Moran et al. 2014; Elming et al. 2014), which include different functionalities to record keyboard data and PE time, sometimes combined with eye tracking data or human evaluations. The key goal of gathering data for measuring PE effort lies in finding ways to potentially *decrease* the effort. Identifying how specific features of the ST and the MT connect to PE effort could help to create technical ways to reduce the effort, as well as to predict the usability of MT. As Krings (2001, 181–182) points out, cases where PE effort is particularly high or particularly low are of special interest, and detailed investigations of the features involved would serve as bases for making decisions.

2.3.2 Machine translation errors and effort

As noted before, PE effort is naturally much affected by the MT quality: deficiencies or errors in the machine translated text need to be corrected, which leads to effort. The term *error*, itself, is not unproblematic. How are errors defined, exactly, and who decides what is or is not an error? As there generally are more than one possible way of translating any given sentence, it is not always easy to pinpoint precisely where the error lies. For the purposes of this study, MT errors are identified and defined through the actions of the post-editor, and the edits made are seen as reflections of errors. For example, if the post-editor changes the order of certain words in a sentence, that edit indicates a word order error in the MT. If the post-editor replaces one word with another, that reflects a mistranslated word in the MT. By comparing the words in question, we can further classify the error in question, for example, whether the inserted word is of the same word class (noun, verb, etc.) as the deleted one.

Although the sheer number of errors in MT certainly has an effect on PE effort, not all errors are equal. Authors like Bensoussan and Rosenhouse (1990) have shown that some errors are more critical to the meaning of a sentence as a whole, and translation quality evaluation models generally distinguish different levels of error severity, such as *minor*, *major*, and *critical* errors (see O'Brien 2012, 62-63). Distinctions of minor and major errors in MT also appear already in early writings based on post-editors' practical experiences in MT. For example, Green (1982, 101) describes minor errors as those involving missing definite articles, incorrect prepositions, personal pronouns, or mistranslated nouns, while major errors involve literally translated idioms, part-of-speech errors, or active versus passive verb forms. The question of why different MT errors would require different amounts of PE effort is also brought up by Krings (2001). Some errors may be easy to detect and only require a few operations while others may require both more intensive cognitive processing and more extensive rewriting. Even more importantly, the amount of rewriting does not necessarily equal the amount of effort needed to detect the error and decide how to correct it. Some errors may be easy to detect but involve several technical operations to be corrected, while others may require considerable cognitive effort although the correction requires only a few technical operations. According to Krings (2001, 179), in such cases the amount of cognitive effort should be seen as the decisive factor.

To examine this question, an error typology specifically for classifying MT errors in terms of cognitive effort was suggested by Temnikova (2010). This classification defines 10 error types, building upon an earlier MT error classification (Vilar et al. 2006), and ranks the error types in accordance to presumed cognitive PE effort. The ranking is based on research related to the cognitive model of reading, working memory, and written language error detection as well as PE experiments. The classification includes the following error types, ranked from easiest to most difficult: 1) incorrect form of correct word (morphological error), 2) incorrect style synonym, 3) incorrect word, 4) extra word, 5) missing word, 6) erroneously translated idiomatic expressions, 7) wrong punctuation, 8) missing punctuation, 9) word order error at word level, and 10) word order error at the phrase level (Temnikova 2010, 3488). This ranking is then investigated using data from a previous PE experiment. Temnikova (2010, 3489–3490) reports that fewer errors of the types presumed to be more cognitively difficult were found in simplified texts, which had previously been found to involve less effort (as measured by PE time and edit distance), than in more complex texts. It should to be noted, however, that cognitive effort may not be the only aspect making morphological errors easier than word order errors, as both PE time and edit distance also reflect the technical effort involved.

A slightly different categorization is used by Lacruz et al. (2014) with the objective of providing "a simple cognitively-based classification of MT errors". The classification is based on a grading rubric used in translator certification (Koby and Champe 2013). The classification suggested by Lacruz et al. contains five error types: mistranslation, omission or addition, syntax, word form, and punctuation (2014, 77). This is, however, not offered as a simple ranking of errors from most to least demanding. The distinction of *mechanical* errors and *transfer* errors should also be considered. Mechanical errors are defined as errors that do not affect the meaning or usefulness of the TT, and can be readily recognized without checking the ST, whereas transfer errors affect the meaning (Koby and Champe 2013, 166). With regard to PEMT, Lacruz et al. (2014, 76–77) hypothesize that transfer errors involve higher cognitive demand than mechanical errors, but point out that errors in each of the categories suggested can be either mechanical or transfer errors, depending on the context. This categorization is investigated using pause data, ratings given by post-editors, and edit distance as indicators of effort. The results reported show that increased effort measures did in fact correlate most strongly with errors classified as transfer errors, and errors categorized as mistranslations and omissions or additions (Lacruz et al. 2014, 81–82).

PE effort related to specific MT errors and edit operations was also explored by Popović et al. (2014). They focussed on five types of edits – word form edits, word substitutions, additions, omissions, and reordering of words – and analyzed discrepancies found using PE time information to measure temporal effort, and human evaluations to measure cognitive effort. The analysis suggests that the cognitive effort, in terms of human evaluation, appears mainly affected by reordering and mistranslated words, and correcting mistranslated words took the longest time in post-editing (Popović et al. 2014, 197).

While these studies have offered some indications of how the type of MT errors, in addition to number of errors, affects PE effort, more work remains to be done. For example, studies so far have used a rather limited number of source and target languages, and it is not entirely clear whether the categorizations are generalizable across languages.

2.3.3 Source text features and post-editing effort

Another question related to MT errors and PE effort is whether specific features of the ST are associated with potentially increased PE effort. Certain features of the ST have been observed to be problematic for MT systems, affecting the MT quality and thereby the effort involved in PEMT. Such features have been referred to as *translatability indicators* (Underwood and Jongejan 2001, 363), or *negative translatability indicators* (O'Brien 2005, 38). The term used by O'Brien (2005) is more descriptive, in that the features specifically affect translatability in a negative way. Most of the research involving these features has focused on English, and they may be to some extent dependent on the languages and MT systems in question, although Bernth and Gdaniec (2001)[2] argue that the overall principles are probably applicable across languages.

Commonly cited features affecting machine translatability involve the sentence length and sentence structure. Both very long and very short sentences can be problematic (for example, Bernth and McCord 2000; Bernth and Gdaniec 2001; Underwood and Jongejan 2001): long sentences can be hard to parse, and short sentences may be ambiguous due to the limited context. In terms of sentence structure, incomplete sentences where the predicate verb or obligatory arguments of the predicate are missing, and the presence of coordinating conjunctions have also been suggested to be problematic for MT (Bernth and McCord 2000; Underwood and Jongejan 2001). The effect of these features has also been observed in PE studies. For example, when investigating PE effort by comparing PE time to the number of edits, studies have found that both very long and very short sentences tend to take longer to edit than would be predicted by the number of edits made (Tatsumi 2009; Tatsumi and Roturier 2010). Tatsumi and Roturier (2010, 47–49) also found that incomplete sentences as well as complex and compound sentences were associated with longer PE times.

Other ST features negatively affecting machine translatability involve combinations of certain potentially ambiguous parts-of-speech, such as proper nouns, noun compounds, non-finite verbs; or phrase categories, such as prepositional

[2]Article IV incorrectly lists the publication year as 2002.

phrases or long noun phrases; and particularly their combinations, such as multiple consecutive noun phrases or prepositional phrases; or missing punctuation (Bernth and McCord 2000; Underwood and Jongejan 2001). To date, there appears to have been less research on such more detailed features and their relationship to PE effort. Using pause data and CNA to investigate certain negative translatability indicators, O'Brien (2005) found some connection between increased PE effort and a long noun phrase, but the other features examined, like proper nouns and punctuation problems, did not show increased cognitive effort. More recently, Vieira (2014) found some connection between increased cognitive effort and prepositional phrases, as well as instances where the same words appear repeatedly within a sentence.

2.4 Differences between post-editors

In most PE studies involving multiple post-editors, considerable variation has been observed between them with regard to PE speed as well as other factors (e.g. Krings 2001; Plitt and Masselot 2010; Sousa et al. 2011). This observation is hardly surprising. As Krings (2001, 173) points out, translation process research in general has found great individual variation in nearly all the characteristics of translators, and in general, the same is likely to hold for nearly any activity that humans are involved in. Perhaps the most noticeable variation between post-editors seems to relate to PE speed. For example, investigating the differences more closely, Tatsumi and Roturier (2010) found that the post-editors in their study differed more in terms of PE time than the number of edits they made.

Some of the possible reasons for different PE speeds include subject area knowledge, general word processing skills, as well as experience with PE and the tools used (see Koby 2001, 20). It may seem intuitive that more experience would lead to faster PE speeds. In light of the PE studies, however, the situation is not necessarily so simple. For example, Krings (2001, 550) found that experienced translators were in fact slower when post-editing than less experienced ones. In a more recent study, Guerberof Arenas (2014b) found that although translators with little or no experience with PE were the slowest, there was no significant difference between experienced translators and novices. Similar findings were reported by de Almeida (2013, 199). Aranberri et al. (2014) also compared professional translators who had no specific PE experience and lay users who had some expertise in the domain of one of the texts in question. Their results showed that overall, both the translators and the users benefited from the use of MT, but the non-professional translators benefited more, and particularly when the text was in their area of expertise (Aranberri et al. 2014, 31).

The study by de Almeida (2013) investigated experience and "PE performance", which takes into account the post-editors' adherence to PE guidelines, and the fitness for purpose and quality of the post-edited translation in addition to productivity. In her study, de Almeida examined quality through an analysis of the PE changes classified as *essential* or *preferential*, depending on whether or not the change was required to make the MT sentence grammatical and accurate, and took into account also cases where the post-editor had not made an essential change or had introduced new errors (2013, 95). The post-editors who performed the best according to this definition both had previous translation and PE experience (de Almeida 2013, 200–201).

With respect to experience and speed, PEMT appears similar to translation. Studies comparing professional translators and non-professionals or semi-professionals (such as translator students) have not found significant differences in translation speed, either. A potential reason for this suggested by Jääskeläinen (1999, 118) is that non-professional translators may not be as aware of potential translation problems and may therefore proceed faster (see also Englund Dimitrova 2005, 21). A similar hypothesis is suggested by Krings (2001, 550). It is also important to remember that, as discussed in sub-section 2.2.1, translation and PE are different tasks involving at least to some extent different processes. Therefore, experience with translation does not equate to experience with PEMT. Furthermore, as de Almeida (2013, 195–196) notes, individual characteristics play a part in PE performance, and the relationship between performance and experience is complex.

Although speed may be the most evident difference, also other kinds of variations have been observed between post-editors. Krings also discusses individual variation related to revising the MT. According to Krings (2001, 530), some of the post-editors considered multiple alternatives in their TA verbalizations before writing while others did not, and some revised their translations after writing them down while others did not. Krings further notes that there was not necessarily any relationship between the tendency to consider alternatives and plan before writing, and the tendency to revise. One post-editor might tend to plan before writing and also tend to revise, another might tend to do little explicit planning and also little revising, or any other combination of the two. (Krings 2001, 530) In translation, similar observations related to planning and revising have been made by Englund Dimitrova (2005, 152–153), who compares her findings to Krings's, and describes five different translator profiles differing in the relative lengths of the planning, writing and revision phases, and the amount of revising done. de Almeida (2013, 180) also observed different tendencies concerning whether the post-editors revised their translation as the last stage, but found no correlation with experience.

Chapter 3
Research design, data and methodology

This section describes the data and methodology used in this dissertation, as well as the way the research was conducted in practice. The research project consists of various interconnected studies, which will be described in detail below. Each study involved analyses of PE data to explore different aspects of the research questions discussed in 1.2. The studies, as well as the materials and methods are described next. Parts of the research project were conducted in collaboration with researchers in other universities in Finland and the United Kingdom. For these joint research projects and co-authored articles, the division of labor is described in a separate appendix.

3.1 Survey article

Article I presents an overview of the literature related to the history and current state of PEMT, as well research into the productivity and quality of PEMT. Specific focus is given to research related to PE effort, and what is known so far. The survey serves as background for, as well as a synthesis of, the studies forming part of this dissertation. Although this article is chronologically the last one to be published, the survey reported in it reflects work that has been ongoing since the beginning of this dissertation project.

This article is published as part of a special journal issue comprising selected articles from the 7th EST Congress panel "The translation profession: Centers and peripheries".[1] One of the central themes of the panel and the subsequent journal issue involves the relationship between humans and machines in the translation field, which this survey article explores in the context of PEMT.

3.2 Study 1

This section describes the study reported in Article II. The aim of this study was to examine whether the human evaluations of PE effort are linked to the amount of editing performed, or whether specific types of errors and edits can be found in cases where the evaluation indicates a given sentence was particularly difficult or particularly easy (see RQ1a).

3.2.1 Data analyzed in the study

The study reported in Article II is based on post-editing data provided by the organizers of the NAACL 2012 Seventh Workshop on Statistical Machine Translation WMT12 (Callison-Burch et al. 2012). The data analyzed was taken from the training dataset intended for the quality estimation task organized as part of

[1] http://www.fb06.uni-mainz.de/est/62.php

the workshop. This dataset was selected for Study 1 because it offers the possibility of comparing different dimensions of effort: the amount of post-editing visible in the changes made, and the post-editors' perceptions of the effort required as expressed by the evaluation scores.

The dataset contains 1832 segments (generally sentences) of English-Spanish machine translations of news texts, produced by a phrase-based statistical MT system. The data is presented as segments, and includes the English source sentence, a Spanish reference translation produced by a human translator, a machine translation into Spanish, a post-edited version of the MT, and an evaluation score that reflects the amount of post-editing needed to produce a publishable translation. The workshop organizers report that the evaluation of PE effort was conducted by three professional post-editors who evaluated the effort required to edit each sentence on a scale from 1 (cannot be edited, needs to be translated from scratch) to 5 (little or no editing required) (Callison-Burch et al. 2012, 25).

3.2.2 Analysis methods

The data were analyzed to compare the amount of post-editing on each sentence to the human evaluation score reflecting effort. To measure the amount of editing, the edit distance metric (H)TER [2] (Snover et al. 2006) was used to calculate the number of changes between the MT sentences and their post-edited versions (for a more detailed description of the metric, see section 2.3).

In order to focus on cases where the perceived effort score and the amount of editing differed, two types of sentences were selected. In the first case, the sentences selected had a poor evaluation score, indicating much effort was required, but the HTER score showed few edits. These sentences represent cases where the human evaluators perceived the level of effort to be higher than the relative number of changes. For comparison, a set of sentences with similar HTER score but a good human score (indicating little effort) was selected. In the second case, the sentences selected had a good evaluation score but the HTER score showed a large part of the sentence had been edited. These sentences represent cases where the human evaluators perceived the level of effort to be lower than the relative number of changes. For comparison, a set of sentences with a similar low HTER but poor evaluation scores (indicating much effort) was also selected. The total number of sentences analyzed was 144.

To get a more detailed idea of the types of edits in the selected sentences, they were then tagged with the FreeLing Spanish tagger (Padró et al. 2010), which provides the surface form of the word, lemma, and a tag with part-of-speech

[2] In Article II this metric is referred to as TER. However, the common usage in the context of PEMT studies is to use the name HTER, and this form is used in the other articles in the dissertation. Therefore, this summary will also use the form HTER. The metric described is the same in all cases.

(POS) and other grammatical information. This information was first used to align the MT and PE sentences so that edits involving the word forms as well as reordering could be tracked. The statistical MT alignment table provided with the dataset was used to match substitutions that involved a different word. Words appearing in the PE version but not in the MT were labelled as insertions and words appearing in the MT but not in the PE version as deletions.

Based on these alignments, the edits in each sentence analyzed were labelled as word form changes, word substitutions involving the same POS (for example, a noun replaced another noun) or a different POS (for example, a noun replaced by a verb), insertions or deletions. In cases where the word order did not match, the moved word was labelled with the distance it had been moved and whether it had been moved alone or as a part of a larger group of words. The totals of changes within a sentence were then calculated, and the types of edits found in the sample sentences were compared to the comparison sets of sentences with similar HTER scores. Additionally, Spearman rank correlations between the manual effort score and the various edit categories were calculated for all tokens and specific POS classes.

3.3 Study 2

This section describes the study reported in Article III. This study had two goals. Firstly, the aim was to examine whether specific types of errors and edits can be found in cases where the PE effort is increased, compared to cases that can be corrected with relatively little effort (RQ1). For this purpose, effort was approached through a different measure than in Study 1. This study utilized information about PE time and compared sentences with long versus short editing times relative to the length of the sentence (RQ1b). Secondly, the goal of this study was to examine different post-editors (RQ3) by comparing their PE process data (PE time, keylogging data and edits made) on the same set of MT sentences (RQ3a).

3.3.1 Data analyzed in the study

The data analyzed in this study was based on the WMT11 workshop dataset of English-Spanish newspaper texts and MT sentences (Callison-Burch et al. 2011). From this dataset, 299 source sentences were randomly selected, providing altogether 1484 translations. The MT versions came from eight MT systems. The machine translations were then edited by eight native Spanish speaking post-editors, who either were professional translators (six cases) or had some experience with post-editing (two cases).

To investigate the two research questions, the post-edited sentences were divided into two sets. To examine the question of the types of edits found in sentences with long versus short PE times (RQ1b), some of the MT sentences were randomly distributed amongst the post-editors. These consisted of different MT versions of the same 279 source sentences, but each person only edited one translation for a given source sentence. To examine the potential differences between post-editors (RQ3a), a smaller set of 20 source sentences with one MT version were given to all eight post-editors.

The post-editing was carried out using the PE tool PET (Aziz et al. 2012), which collects various types of process data. PE time (total time for each sentence and total divided by the number of words) was recorded for each sentence. To examine the typing done by editors, the number of keystrokes pressed per sentence edited as well as the number of keystrokes divided by the PE time and the type of key pressed was logged. Finally, the HTER score (Snover et al. 2006) was used to measure the relative number of edits performed on a sentence.

3.3.2 Analysis methods

To examine the question of the types of errors and edits found in sentences with long edit times (RQ1b), the PE times of sentences were compared and a set of sentences with long editing times in seconds per word were selected for a more detailed analysis. The time comparisons were made for each post-editor separately, as different editors may have different PE speeds. Time was measured as seconds per word instead of total PE time to avoid only selecting very long sentences that may take a long time to edit due to their length. Sentences with very high HTER scores indicating extensive rewriting were also excluded in order to be able to focus on more specific errors. For comparison, a set of sentences of similar length and with a similar number of changes but with short PE times was also selected. The total number of sentences analyzed was 64.

The sentences selected were then analyzed manually for the types of errors corrected. This error analysis was conducted using an error categorization with ten categories ranked based on the presumed cognitive effort involved in correcting each type of error. The categorization was slightly modified from Temnikova (2010), and covered (from easiest to most difficult): typographical edits, changes to word form, word substitutions involving synonyms, other word substitutions (further categorized according to whether the substitution involved the same POS, a different POS or an untranslated word), extra words, missing words, mistranslated idioms, incorrect punctuation, missing punctuation, word order changes at the word level, and word order changes at the phrase level. The errors found in the selected sentences with long PE times were then compared to errors found in those with short PE times to examine potential differences. To observe edits related to word form changes, the sentences were tagged with the the FreeLing software Padró et al. (2010) for lemma and POS information. All edits were manually labelled according to the error categories used.

To investigate the differences between post-editors (RQ3a), the 20 cases where all post-editors edited the same MT version of the same sentence were analyzed by using the logs stored by the editing tool PET to observe the edits. PE time was used to compare how long the editors took to edit. HTER was used to compare how much they edited in each sentence. Keystroke counts were used to examine how much typing they did to perform their edits. Keystrokes were examined both as a total and according to the type of key pressed. The types of keystrokes were categorized as white keys (space, tab and enter), alphanumeric keys (letters and digits), or control keys (delete, backspace, combinations such as ctrl+c). These categories were used to provide more information about what kind of typing the editors were doing.

3.4 Study 3

This section describes the study reported in Article IV. The aim of this study was to examine the types of errors and edits involving particularly high effort (RQ1) by connecting the edits to specific ST features and using PE time as a measure of effort (RQ1b).

3.4.1 Data analyzed in the study

The data analyzed in this study comes from the set of PE data collected by Elming et al. (2014). This dataset consists of 25 English newspaper documents and their MT versions produced by a statistical MT system from the WMT12 workshop data (Callison-Burch et al. 2012). The texts were translated or post-edited into Spanish by 5 translators during field trials of the CASMACAT workbench (Elming et al. 2014). For the purposes of this experiment, only the PE data were used, resulting in 5-9 documents per editor with a total of 622 sentences. This dataset was chosen for the study because it provides information for analyzing the edits connected to specific points in the source text and for using PE time as a measure of effort.

The dataset includes the source texts, MT versions and one or more PE versions of each sentence. Additionally, the logs contain tokenization and source-target alignments as well as information about the edits. Below the sentence level, the edits have been grouped into "production units", which are sequences of successive keystrokes that produce coherent passages separated by pauses of defined length (Carl and Kay 2011). These units are further mapped to the source words aligned with the edited target words. For each unit, information is included about the duration of the edit in seconds, length of the pause before the next unit, and total numbers of inserted and deleted characters.

3.4.2 Analysis methods

The PE data were analyzed to investigate whether the time spent on editing the sub-sentence level production units could be linked to specific ST features. For this analysis, different types of potential features were defined. The first set of features relates to the potential overall complexity of the whole sentence: the number of words, the number of different phrases (for example, noun phrases consisting of a head noun and modifiers), and the number of predicate verbs and their arguments. The second set of features relates to the type of sentence element being edited (for example, verbs or verb phrases), or their combinations (for example, sequences of noun phrases). For verbs, more specific information about the verb type (finite vs non-finite) was also examined. Named entities, such as names of people or organisations, were also used as a feature. The sentences analyzed were first tagged with SENNA (Collobert et al. 2011) to obtain information about POS, named entities, chunks (phrases), and semantic role labels.

These features were then analyzed against the PE data, specifically the PE time for each unit but also taking into account the number of characters edited. The time and character data were analyzed both as total numbers and relative to the average numbers of seconds or characters for all the production units in a given document by a given editor. For sentence-level analysis, the sentence-level HTER score (Snover et al. 2006) was also calculated. The relationships between the source-text features and indicators of PE time and edits were then analyzed using Principal Component Analysis (PCA) (Jolliffe 2002), which is a technique used to visualise data in text, speech and image processing. This method was used to investigate the patterns of source features correlated with PE time and number of edits.

3.5 Study 4

This section describes the study reported in Article V. The aim of this study was to investigate cognitive effort from the perspective of identifying and correcting MT errors without the help of the source text (RQ2).

3.5.1 Data analyzed in the study

The dataset analyzed in this study consists of of two English newspaper articles with 28 and 32 sentences, respectively, MT versions into Finnish produced by two MT systems (one statistical and one rule-based), and PE versions created by 48 post-editors. The post-editors were native Finnish speaking translator students, and the PE data were collected during a university course on translation technology. Each post-editor edited one article according to their understanding, and they were also given the option of marking each sentence as either acceptable without corrections or as too defective to correct. The data was chosen for this study because it offers a chance to investigate cognitive effort from the point of view of identifying MT errors and being able to correct them.

3.5.2 Analysis methods

To investigate to what extent the post-editors were able to correct the MT errors without the ST, both the raw MT sentences and the PE versions were evaluated for correctness of meaning and language. Meaning and language were evaluated separately at the sentence level, forming four categories: sentences that have both correct meaning and correct language, sentences that convey the meaning correctly despite language errors, sentences that have correct language but the meaning is not conveyed correctly, and sentences that contain errors related to both language and meaning.

Furthermore, sentences where post-editing appeared to have been easy and cases where post-editing appeared to have been particularly difficult were examined. Cases where all (or nearly all) post-editors had successfully corrected a sentence, as well as cases where all or nearly all post-editors had failed to correct the sentence were identified. These example sentences were then analyzed more closely to see whether specific types of errors had created particular difficulties.

3.6 Study 5

This section describes the study reported in Article VI. The aim of this study was to examine the differences between post-editors in terms of their post-editing choices (RQ3b). The study approaches this question in two ways, by comparing which MT version of the same source sentence was selected, and examining the number of different versions created by post-editors.

3.6.1 Data analyzed in the study

The data analyzed for this study come from the dataset created during the evaluation of a multilingual, controlled language text generation and machine translation system (Rautio and Koponen 2013). The dataset consists of 139 English source sentences, three Finnish MT versions of each sentence, and post-editing data from 11 post-editors. The sentences in the dataset are short and structurally relatively simple “tourist phrases”, which involve asking directions, buying things, and small talk questions. This dataset was chosen for this study because it offers the chance to investigate sentences containing relatively few errors, thereby making it easier to compare choices made by several post-editors.

The post-editors were native Finnish speaking translator students at the University of Helsinki. The PE data were collected using the online MT evaluation tool Appraise (Federmann 2012). The post-editors were shown the 139 source sentences and the three MT versions in random order. Each post-editor first selected which MT version they considered best, and could then either accept it without edits or post-edit as necessary. In addition to the source, MT and PE versions, the data also contains PE time data.

3.6.2 Analysis methods

The data described above were then analyzed to see whether the same MT version was selected by all 11 post-editors, and whether they accepted or edited the MT. All the sentences were categorized according to these observations, and the PE versions created by the post-editors were compared to calculate the number of different versions created for each source sentence. The amount of post-editing performed by each post-editor was calculated using HTER (Snover et al. 2006), and the post-editors' scores were compared. PE time was compared using the time data recorded by the evaluation tool. The post-editors were also compared in terms of how often their choice of the best MT version differed from the most commonly selected one, and whether their final PE version differed from the most common translation.

To examine why specific MT versions were preferred by the post-editors, the selected and rejected MT suggestions were assessed for correctness of meaning and language. To investigate how the post-editors' final versions differed, cases with multiple PE versions were compared. The situations where different versions were created and the differences in the PE versions were also examined.

Chapter 4
Results

This section presents the main results of the original articles as they relate to each of the three Research Questions discussed in Section 1.2.

4.1 MT errors, ST features and effort

RQ1 What types of errors and edits can be identified in sentences requiring much effort or little effort?

(a) Using human evaluations as an indicator of effort?
(b) Using post-editing time as an indicator of effort?

Based on the literature concerning PEMT and PE effort, surveyed in Article I, research in general has identified features which appear connected to increased effort. Sentence length and structural features related to sentence complexity have been observed by various researchers to be connected to PE effort. Certain specific ST features like noun sequences and prepositional phrases have also been identified as features increasing effort. In terms of MT errors, edits involving word order and correction of mistranslated idioms are commonly observed to be connected to increased effort. (Koponen 2016) RQ1 is further addressed by Article II, Article III and Article IV.

The study reported in Article II relates to sub-question RQ1a. This study involved an analysis of edits performed in sentences categorized as requiring much or little effort based on human evaluations of the PE effort involved. The analysis focused on two types of cases where the human evaluation of effort required differed from the amount of editing visible in textual changes (Koponen 2012). Firstly, the study examined cases where human evaluators indicated that much effort was needed but the number of words edited was low relative to the total number of words. Secondly, the study examined cases where the human evaluators indicated little editing was needed but the number of words edited was high relative to the total number of words. In both cases, the sentences were compared to cases with a similar number of edits relative to the number of words, but opposite human evaluation scores.

The results from these comparisons show that sentence length tended to affect the human evaluation of effort: in both cases, sentences that were evaluated to involve much effort were longer than those evaluated to involve little effort (Koponen 2012, 186). Because the sentences are long, they also tend to contain a large absolute number of errors, although the number of edited words may be low relative to the total number of words in the sentence being edited. The length of the sentence may also require more effort from the post-editor to identify the defective passages. Sentences evaluated to involve much effort also tend to contain more errors related to word order and word substitutions – particularly substitutions with a different POS – while sentences evaluated to involve less effort contained more edits of the word form. When examining specific word classes, it also appears that sentences evaluated to involve much effort contained more edits to verbs, for example, and those evaluated to involve less effort contained more edits involving determiners, or specific cases such as changing the form of adjectives (Koponen 2012, 187–188).

The study reported in Article III relates to sub-question RQ1b. This study involved an analysis where sentences identified as having comparatively long PE times were analyzed for the types of MT errors and edits involved, and the edits were compared to sentences with relatively short PE times. The edits identified in the sentences were labelled using an error categorization ranked according to the presumed cognitive difficulty of editing each type of error (based on Temnikova 2010).

The results of the error analysis showed some differences between the errors in sentences with long and short PE times. In both types of sentences, most common edits were substituted words, but sentences with long PE times contained more substitutions that involved changing the POS or correcting an untranslated source word. Sentences with long PE times also contained slightly more errors related to idiomatic expressions and word order (especially cases where the reordering crossed phrase boundaries), as well as missing punctuation. Sentences with short PE times, on the other hand, contained more edits involving word substitutions that could be considered synonyms, edits to the word form, and incorrect punctuation. Missing words appeared to be more common in sentences with long PE times. Extra words, particularly extra function words such as determiners, were more common in those with short PE times. (Koponen et al. 2012)

The study reported in Article IV also relates to sub-question RQ1b. This study also utilized PE time as an indicator of effort, but approached the question of edits and PE times from a slightly different perspective. In this study, the edits were connected to specific ST features that may cause particular difficulties in MT, and these features were then examined in terms of the PE time.

The findings of this study show again that PE time is not only connected to the relative number of edits within a sentence. Sentence length again played a role, with long PE times often found in long sentences. However, on the sub-sentence level this correlation was not strong and was mostly observed in long sentences that overall required relatively little editing. With regard to specific features, some connection was observed between longer PE times and units involving verbs, particularly modal verbs, and sequences of consecutive noun phrases. (Aziz et al. 2014, 196)

To summarize the results related to RQ1, the findings of the studies discussed above and the literature survey suggest that the relative number of word-level edits is not sufficient alone to explain PE effort (Koponen 2016). One factor affecting effort, both in terms of human evaluation of effort and the time aspect of effort, is the length of the sentence being edited: long sentences tend to involve more effort even when they contain relatively few edits (Koponen 2012; Koponen et al. 2012). This connection becomes less clear on the sub-sentence level, meaning that time consuming edits as such do not necessarily appear in long sentences (Aziz et al. 2014). Different types of errors and edits may also involve different levels of effort. Edits involving the correction of word form appear to involve less effort than word substitutions, especially if the substitution involves changing POS. Reordering words also appears to involve increased effort, particularly when the words need to be moved across phrase boundaries. Inserting missing words appear to involve more effort than deleting extra words, particularly if these extra words are function words. More effort may also be involved in correcting specific POS or sequences, such as verbs or sequences of consecutive nouns. (Koponen 2012; Koponen et al. 2012)

4.2 MT errors and cognitive effort in PE without ST

RQ2 How do MT errors affect cognitive effort when post-editing without a source text?

Based on Article I, literature related to PE without ST shows that some level of success in correcting the MT output has often been observed. However, relatively little work appears to have been done on analyzing the specific types of MT errors and how successfully they can be edited. Mainly the analysis of PE without ST appears to focus on the sentence-level correctness or acceptability of the post-edits, or comparing quality evaluations (fluency, accuracy, comprehensibility) of sentences post-edited with and without the source text (Koponen 2016). RQ2 is further addressed by Article V.

The question of how successfully specific types of errors can be edited was examined in Article V through an experiment involving PE without source text. The results of this study show that, overall, the post-editors were able to arrive at the correct meaning in about half of the sentences edited. However, a relatively large number of these post-edited sentences still contained language errors, such as grammatical errors that were left uncorrected, or in some cases even new errors (for example, typographical errors) introduced by the post-editors. Nearly a quarter of the sentences (331 out of 1444) were left unedited by the post-editors and labelled as “unintelligible”, meaning that they did not feel able to even attempt correction. (Koponen and Salmi 2015, 123) The results also showed that in many cases, the success of post-editing varied: some post-editors were able to correctly post-edit a sentence while others were not, or did not even attempt to make corrections. Some sentences, however, could be identified as particularly easy, based on all or all but one of the post-editors succeeding to correct the meaning. Other sentences could be identified as particularly difficult, based on none or at most one being able to post-edit the sentence correctly. (Koponen and Salmi 2015, 127)

These results can be seen to address PE effort in two ways. Firstly, the sentences labelled as too defective to even attempt correction provide information about the situations where the post-editors themselves consider the effort to be too high. A closer analysis of these sentences showed that, in general, these cases involve long sentences with multiple errors. However, there were also cases where even one error can significantly change or obscure the meaning, making post-editing impossible. Secondly, the analysis of the types of errors in sentences that were particularly easy or particularly difficult to edit provides information about the cognitive effort involved in identifying and correcting errors. The easily-edited sentences show that certain types of errors can be easily identified and corrected even when the ST is not present. These were generally MT errors involving the word form of an otherwise correct word, if the relations between words in the sentence could be deduced based on context and general knowledge (Koponen and Salmi 2015, 128). On the other hand, it was also observed that in cases where the incorrect word forms interfere with the interpretation of the syntactic relations, post-editing could turn out to be impossible. Analysis of the sentences that were particularly difficult to correct often involved mistranslated words and idioms, or missing information (Koponen and Salmi 2015, 129–130). These types of errors may then be generally more difficult to identify and correct.

4.3 Differences between post-editors

RQ3 How do different post-editors and their edits differ from each other?

(a) In terms of effort indicators related to process data?
(b) In terms of preferences related to editing choices?

Based on the literature surveyed in Article I, research has generally shown variation between post-editors in various ways. Post-editing speeds and productivity vary between post-editors, as does the change in productivity when compared to translation from scratch. While increases in productivity are generally observed in the research literature, the level of increase varies from one post-editor to another. General observations have also been made with regard to the amount of editing and the methods of post-editing (Koponen 2016). RQ3 is further addressed by Articles Article III and Article VI.

The study reported in Article III explores sub-question RQ3a by analyzing different post-editors editing the same MT sentences. Their process data was recorded and PE time, edit distance and keylogging data was compared between post-editors. The comparison of these cases showed that post-editors differ from each other in various ways. Firstly, comparing the number of edits in a sentence showed that even with the same instructions to minimally change the machine translations, the number of changes made differed. Secondly, comparing the keylogging data showed that post-editors make these changes in different ways. Some make use of as many MT words as possible, editing only those characters that needed to be changed and using cut-and-paste operations for reordering, while others delete the MT words and rewrite their own version even when many of the words in fact are the same. Thirdly, post-editors appeared to differ in how they approached the planning and typing of corrections. Some planned the corrections first and made them in one round of consecutive changes within the sentence, while others revised their own corrections, moving back and forth in the sentence. (Koponen et al. 2012) The comparisons also showed that PE time and the amount of editing undertaken do not necessarily correlate with each other, and post-editors may make a similar number of edits in varying amounts of time.

The study reported in Article VI addressed both sub-questions RQ3a and RQ3b. This study involved an analysis of post-editors selecting their preferred MT version out of three options and editing as necessary. One part of the analysis compared the PE time and edit distance of different post-editors (RQ3a). The other part examined the number of different versions created by post-editors as well as the tendency of each post-editor to produce versions varying from the most common version (RQ3b).

With regard to sub-question RQ3a, the results from the comparison of PE times and edit distances showed that there were relatively few differences in the amount of editing done by different post-editors. Some variation appeared in the number of sentences they chose to post-edit, as opposed to accepting them without corrections. PE times showed more variation, and a comparison of edit distances to PE time showed that the number of edits and PE time do not necessarily correlate. Rather, certain differing profiles were observed, where some post-editors may edit fast but still make many changes, or conversely, edit slowly but make few changes. (Koponen 2013, 6–7)

With regard to sub-question RQ3b, the results of comparing the MT versions preferred by the post-editors and the final translations they created showed that the post-editors were mostly in agreement on which was the best MT version, and in most cases only one or two PE versions were created (Koponen 2013, 4). This small number of versions is likely related to the nature of the controlled language data, which contained very few ambiguous cases. A larger number of PE versions was generally created in situations where a sentence could be considered to be ambiguous. The analysis of how the post-edited versions differed showed that at least some post-editors had strong preferences for particular wording even when the substituted words were synonymous. (Koponen 2013, 7–8)

To summarize the results related to RQ3, the findings of the studies forming part of this dissertation and the literature survey suggest that post-editors differ in various ways. Different types of post-editor profiles can be observed, based on how much each post-editor edits, how fast they edit, and how they approach the post-editing (Koponen et al. 2012; Koponen 2013). Post-editors differ in whether they plan their edits before making them in one round or revise their own corrections, and in whether they maximize the use of the MT or prefer to write their own versions (Koponen et al. 2012).

Chapter 5
Discussion

This chapter discusses the results reported in Chapter 4 in the wider context of the theoretical and practical context presented in Chapter 2. The results of this study contribute to the theoretical understanding of PE effort and its relationship with MT errors and ST features, as well as research on PE processes and how indicators of PE effort vary between different post-editors. The results provide evidence that, in addition to the number of edits performed, post-editing effort is affected by the type of edits as well as source text features. Secondly, the results show that while certain language errors can be corrected even without access to the source text, certain other types that more severely affect the meaning cannot. These findings may suggest to MT researchers and developers potential features to be targeted in order to reduce effort, or suggest to MT users ways to identify source texts or machine translations that are likely to be unsuited for PEMT. Thirdly, the results show that post-editors' speed and amount of editing performed differ, and that various profiles can be identified in terms of how the edits are planned and carried out by the post-editors, which may have both theoretical and practical implications for measurement and estimation of post-editing effort.

The potential theoretical and practical implications are discussed next in more detail. This chapter also presents an evaluation of the research design, and finally some recommendations for future research.

5.1 Theoretical implications

Based on the results of this study, there are certain implications that can be used to improve the measurement of effort in post-editing. A common practice for measuring PE effort involves the use of edit distance metrics such as HTER (Snover et al. 2006), which express the effort in terms of number of edits performed in a sentence relative to the number of words in a sentence. However, the comparisons of these indicators of effort against subjective effort evaluations (Koponen 2012) and PE time (Koponen et al. 2012) presented in this dissertation show that edit distance alone cannot fully capture PE effort. While it does to some extent reflect the effort needed, cases can be observed where post-editing the sentence takes more time (or less) or is perceived as more demanding (or less) than would be expected based on the relative number of edits.

One of the explanations for these discrepancies are the lengths of the sentences. The effect of sentence length was observed particularly in the analysis presented in Article II, where it was shown that human evaluators tend to assess the PE effort related to long sentences as high even when relatively few changes were performed (Koponen 2012, 188). These results support other findings where sentence length was found to affect also the PE time (Tatsumi 2009; Tatsumi and Roturier 2010; Popović et al. 2014). In the earlier studies involving PE time, it has been found that in addition to long sentences, very short sentences also require more time than predicted by the number of words (Tatsumi 2009). This effect was not observed in the comparison of edit distance and the subjective evaluations of effort. In the study reported in Article II, short sentences tended to be evaluated as involving little effort even when the numbers of edits were relatively high (Koponen 2012). This may be explained by the fact that the absolute time when post-editing long sentences is much longer than when post-editing short sentences. Therefore, the human evaluator perceives the effort as low in relation to short sentences, even if the relative amount of time is high. Overall, when edits and PE time were examined at the sub-sentence level in Article IV, a modest correlation was also observed, but it should be noted that time-consuming edits cannot be generalized to necessarily occur in long sentences (Aziz et al. 2014, 196).

In addition to length, other features of the sentence being edited were found to affect PE effort. Sentence structure and complexity have been suggested as source features that may affect effort Tatsumi and Roturier (2010, 47–49). The results reported in Article IV offer some support for this, in that some connection was found between the number of predicates and arguments in a production unit being edited and the PE time (Aziz et al. 2014, 195). More specifically, the results of our study also suggested that longer PE times may be linked to specific features such as verbs or sequences of consecutive nouns (Aziz et al. 2014, 196). These results also support other findings related to ST features that may lead to increased PE effort because they cause difficulties in MT (Bernth and McCord 2000; Underwood and Jongejan 2001), or to the post-editors (O'Brien 2005; Vieira 2014). On the other hand, some specific features that have identified as problematic for MT, for example gerunds and other non-finite verbs, were not observed to be linked to longer PE times in our study. At least based on our results, while these features may lead to MT errors, they do not necessarily lead to increased PE effort (Aziz et al. 2014, 196).

The results of this study suggest that also the type of MT error to be edited affects PE effort. The studies forming part of this dissertation explored this issue by comparing the errors and edits that occur in cases where post-editing appears to be particularly difficult or particularly easy, based on human evaluations of effort (Article II) or PE time (Article III). The results of the study support earlier findings that not all errors and edits are equal, rather, some are more demanding than others. Edits to the form of an otherwise correct word appear to be relatively easy, while edits such as reordering, missing words, and mistranslations, particularly mistranslated idioms appear to be more difficult (Koponen 2012; Koponen et al. 2012). Similar findings have been made in other studies (Temnikova

2010; Popović et al. 2014; Lacruz and Shreve 2014; Lacruz et al. 2014).

While the results of this dissertation mostly support the previously proposed cognitively motivated classification of errors (Temnikova 2010), which was used with slight modifications in the study reported in Article III, the results also contradict some specific details. In Temnikova's classification, punctuation errors (missing or incorrect punctuation) are assumed to be among the more cognitively difficult cases. In our study, little difference was observed between the sentence types. In fact, incorrect (as opposed to missing) punctuation was only found in the sentences identified as easy (Koponen et al. 2012). The cognitive effort caused by punctuation errors therefore seems less clear, although in some situations missing or incorrect punctuation could certainly change or obscure the meaning of a sentence. Similarly, our study found more cases of correcting extra words in sentences identified as easy. Mostly these involved function words, suggesting that at least extra function words do not necessarily involve particular effort. (Koponen et al. 2012) We also found it necessary to modify the classification by dividing the category of "mistranslation" (from Temnikova 2010) into three cases based on whether correcting the mistranslation involved substitution with a word of the same POS, a different POS, or correcting an untranslated source word. Easy sentences were found to contain more cases where the POS remained the same, while changes to POS and untranslated words were more common in difficult sentences (Koponen et al. 2012).

The consideration of whether specific types of errors are cognitively difficult or not may also be less straightforward. This point is made by Lacruz et al. (2014, 76–77) who argue that, depending on context, even errors in the same category can be mechanical errors, which do not affect the meaning and can be easily recognized without the ST, or transfer errors, which do affect the meaning and are not as easy to recognize (for this distinction, see also Koby and Champe 2013, 166). The results of this dissertation show that analyzing the success of PE without ST offers insight into which types of errors can be identified and corrected without the ST. The results reported in Article V support the previously noted observations that word form errors can be easy to post-edit, while missing information and mistranslations of words and idioms can be particularly difficult to recover. However, word form changes were found to be easy only in cases where they did not interfere with understanding of the syntactic relations between the words (Koponen and Salmi 2015, 128–129). The language used in our study was Finnish, where the morphological form of the word carries much of the information about syntactic relations, and depending on the context, even one word form error may significantly change or obscure the meaning. Earlier findings by Krings (2001, 272) have also suggested that especially mistranslations with incorrect POS were difficult to correct without ST. Krings' analysis also showed punctuation errors and word order to be easy to correct. Word order errors were not seen as particularly easy or difficult in our study, which may at least in part be connected to the language of this PE task being Finnish, which has relatively free word order. Punctuation errors are somewhat difficult to assess in our data. Although in many cases our post-editors successfully corrected punctuation, in general it appears that many of them were not particularly attentive to language, leaving errors unchanged or in some cases even introducing new errors (Koponen and Salmi 2015, 131). It is therefore difficult to say whether they were unable to correct, or simply did not notice or consider punctuation important.

The results of this dissertation, particularly Article III and Article VI, support earlier findings that the PE process varies between different post-editors (Krings 2001; Plitt and Masselot 2010; Sousa et al. 2011; Guerberof Arenas 2014b; de Almeida 2013). Similarly to the findings of Tatsumi and Roturier (2010), the studies forming part of this dissertation also found that post-editors differ more in terms of PE time than the number of edits they make (Koponen et al. 2012). The differing pause and typing patterns identified also support earlier findings by Krings (2001) that some post-editors appear to plan their corrections before typing them only once, while others make multiple revisions (for variation in planning profiles, see also Englund Dimitrova 2005, 152–153). PE speed, the number of changes made, amount of typing and revising were also not necessarily connected to each other. Rather, differing profiles emerged. Some post-editors may work fast while also making many changes, while others might be slow but make relatively few edits (Koponen et al. 2012).

With regard to how the edits performed by post-editors differ from each other, the results reported in Article VI showed that at least when relatively short, simple, and high-quality MT sentences are post-edited by multiple post-editors, most sentences tended to have only one or two PE versions (Koponen 2013, 5). This finding is similar to Tatsumi et al. (2012). A closer comparison of the PE versions in this study also showed that at least some differences appeared to be connected to individual preferences regarding specific wording (Koponen 2013, 5). This observation relates to earlier findings by de Almeida (2013, 186) that between 16–25% of PE changes could be categorized as preferential rather than essential for making the MT sentence accurate and grammatical. Taken together, these findings show that care must be taken when edit distance metrics are used as a measure of effort, as they rely heavily on the assumption of the changes made by the post-editors being both correct and necessary.

5.2 Practical implications

In addition to theoretical considerations, the results of this dissertation may have some practical implications. The results indicate that the commonly used approach to measuring effort, edit distance, does not alone accurately capture the actual amount of PE effort. Rather, there may be a need to take into account other features such as sentence length and types of edits. The type of MT error being corrected may affect both technical effort, in that a large number of edit operations are needed to correct it, and cognitive effort, in that identification or correction of the error is not simple. The results suggest that as an indicator of PE effort, keylogging data regarding the amount of typing can capture the technical aspect of effort, and PE time can be a more accurate measure because it combines both the technical and cognitive effort (Koponen et al. 2012).

The variation between different post-editors observed in this study also has implications with regard to the use of various indicators of effort. In addition to the difficulty of measuring cognitive effort, the observations regarding the way post-editors actually perform the work of post-editing also call into question some of the assumptions also about the amount of technical effort involved and whether it is apparent through the edit distance. For example, a word form edit may appear to only require the deletion and addition of a few characters, or reordering could be assumed to occur as one cut-and-paste operation. However, the analysis of post-editors' process data showed that at least not all of them work in this manner. They may instead prefer to delete and rewrite large passages despite the final result being relatively close to the deleted MT suggestion (Koponen et al. 2012). Closer investigations of these types of processes may also offer further insight into the training of post-editors: how different types of edits are performed, and how they can be performed most efficiently. These results are also connected to the practical issue of compensation for PEMT work. The observed differences between post-editors support the argument made by Guerberof Arenas (2014a, 183) that assuming a general level of productivity increase, and compensating the post-editors according to the presumed time savings may not be fair to the post-editors. Further, productivity should not be over-emphasized at the cost of quality.

The results of comparing edits in sentences identified as easy or difficult also suggest some specific types of MT errors and ST features that appear to increase PE effort. These results may have implications in the field of MT development by suggesting which features could be targeted for improving the quality of MT specifically for PEMT purposes. They may also be useful for the field of Quality Estimation (see Specia et al. 2010b, 183), which aims to estimate the quality of a given MT suggestion for a given ST, and could be used either to provide the post-editor with information about the estimated quality of the MT suggestion, or even to filter out suggestions likely to be of little use.

5.3 Reliability and limitations

While the analyses of sentences identified as easy or difficult suggest specific types of features that appear to be linked to much or little PE effort, it needs to be noted that the numbers of sentences analyzed are relatively low. The relatively low number of cases limits the strength of any general conclusions that can be drawn based on these results. Nevertheless, they do offer qualitative results that may be useful as starting points for further quantitative research exploring larger datasets.

The data is also limited in the number of languages involved: the source language in each study is English, with Spanish and Finnish used as target languages. This selection of languages is likely to have affected the results to some extent. Specific ST and TT features discussed in the results may not be generalizable across different types of languages. More extensive studies using different language pairs would naturally be needed to investigate to what extent the same types of features can be linked to PE effort in other languages.

The nature of the corpora used may also affect generalizability. Most of the articles forming part of this dissertation involved the use of newspaper articles. These types of texts are commonly used in the field of MT and PEMT research, and were used in this research project as well, due to the open availability of such corpora (Callison-Burch et al. 2011, 2012; Elming et al. 2014) and to potentially increase the comparability of the results. Another benefit of these texts is that they are fairly general, and therefore post-editing them does not require any specific domain knowledge. The use of newspaper articles, both in this study and in general, does pose some issues with regard to ecological validity in that they differ from the types of texts that are likely to be post-edited in real-life practical scenarios, for example, user manuals or other technical texts. For further studies, the ecological validity of the text type warrants further consideration.

In one of the studies forming part of this dissertation (Article VI) the corpus used involved a rather limited sample of controlled language data, comprising a set of isolated sentences formed by a small number of defined vocabulary and structures. The nature of the data is likely to have affected the results, and a less controlled text type would likely have revealed more variation in the post-editors' choices. However, the repetitive vocabulary and structures as well as the high quality of the MT were useful in offering a chance to observe post-editor choices across similar cases and attempt to isolate their changes to specific features, as relatively few changes were made by the post-editors.

In the case of two of the studies (Article V and Article VI), the post-editors involved were translator students with no prior post-editing experience. From prior research, it is known that the processes and products of translator students differ to some extent from those of professional translators or post-editors (for example, Englund Dimitrova 2005). Care must be taken, therefore, when generalizing any results involving students to translators or post-editors in general.

Overall, it is also important to note that the analyses performed as part of this dissertation generally do not involve evaluation of the post-edited versions. The post-editors' corrections were assumed to be correct and necessary. This is connected to the definition of *error* that this dissertation has operated on – an error is any feature that causes the post-editor to make a change – and the focus on the processes rather than the products of post-editing. Although this definition was deemed useful for the purposes of this study, it relies on the assumption that all changes by post-editors are correct and necessary. This assumption is not wholly unproblematic. The two studies included in this dissertation where a closer analysis of the post-edited versions was conducted (Article V and Article VI) do show that some caution is also necessary. Examining the post-edited versions used in Article V show that the post-editors in this study, at least, had not been particularly attentive to language errors and sometimes even introduced new errors (Koponen and Salmi 2015). This may be related to the fact noted above that these specific post-editors were translator students, and professional post-editors might have produced more polished versions. Similar observations were made by de Almeida (2013, 135–136), who found that the post-editors sometimes failed to make essential changes or introduced new errors.

Another question related to the evaluation of the post-edited versions is whether, strictly speaking, the corrections are necessary. As can be seen from the studies showing variations between the post-editors, the perception of what needs to be corrected, to what extent and in what way differs even among professional translators. A closer look on this issue is provided by de Almeida (2013), whose results show up to 25% of the changes analyzed could be classified as preferential and not essential. Therefore, particularly with regard to cases where only one MT and PE version was analyzed, it is necessary to be cautious about generalizing the edits to all post-editors and all situations. Rather, the results suggests tendencies in what appears to trigger edits.

5.4 Recommendations for further research

As noted in the previous sub-section, the studies so far are based on rather small set of data and limited number of sentences. The analysis of particularly easy and particularly difficult cases, suggested already by Krings (2001), does in general offer potential topics for future study. In future research, more extensive quantitative studies would be useful to further investigate the qualitative feature types suggested by this study. For future research, it would be particularly interesting to investigate PE data containing multiple different indicators of effort to compare the relationship of these indicators and specific edit types, which has been explored recent studies (Popović et al. 2014). A related issue would involve the investigation of mechanical errors versus transfer errors (Lacruz et al. 2014): how to determine which errors affect meaning and which do not, and how the context affects this determination.

Much of the work on PE effort to date has focused on relatively few languages, often with relatively little morphology. As also noted by Popović et al. (2014), it would be of particular interest to investigate whether, and to what extent, the findings and assumptions about the relationship between PE effort and specific types of features hold in languages with different structure and richer morphology. Finnish as a target language would be an interesting case to study, for example, the effect of word order and word form errors on PE effort. Unlike many of the languages investigated so far, such as English and Spanish, Finnish has a relatively free word order, with much of the information about syntactic and semantic relations between words being carried in the morphological form of the word. Therefore, it would be interesting to investigate whether word form errors in Finnish involve more PE effort than in some other languages. As a wider point, it would be interesting to investigate the relationship between lexical errors and errors affecting meaning through interfering with the syntactic and semantic relations between the words, and how these two error types relate to PE effort.

Another potentially fruitful direction of research would be a closer analysis of the variation in edits performed by different post-editors. Studies on a larger scale, comparing post-editors working on the same (or same type of) texts would be necessary to determine whether, or to what extent, it is possible to generalize edits across different post-editors. Comparing the edits could provide insight into what types of edits can be considered corrections of actual MT problems and whether other types of edits may be more connected to individual preferences.

Bibliography

Allen, J. (2003). Post-editing. In Somers, H., editor, *Computers and Translation. A Translator's Guide*, pages 297–317. John Benjamins Publishing Company, Amsterdam/Philadelphia.

ALPAC (1966). Languages and Machines: Computers in Translation and Linguistics. Technical report, Automatic Language Processing Advisory Committee (ALPAC), Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

Alves, F., Pagano, A., Neumann, S., and Steiner, E. (2010). Translation units and grammatical shifts: Towards an integration of product- and process-based translation research. In Shreve, G. M. and Angelone, E., editors, *Translation and Cognition*, pages 109–142. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Aranberri, N., Labaka, G., Diaz de Ilarraza, A., and Sarasola, K. (2014). Comparison of post-editing productivity between professional translators and lay users. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*, pages 20–33.

Aziz, W., Koponen, M., and Specia, L. (2014). Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In O'Brien et al. (2014), pages 170–199.

Aziz, W., Sousa, S. C. M., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In *8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Bensoussan, M. and Rosenhouse, J. (1990). Evaluating student translations by discourse analysis. *Babel*, 36(2):65–84.

Bernth, A. and Gdaniec, C. (2001). MTranslatability. *Machine Translation*, 16(3):175–218.

Bernth, A. and McCord, M. C. (2000). The Effect of Source Analysis on Translation Confidence. In *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas*, pages 89–99.

Blain, F., Senellart, J., Schwenk, H., Plitt, M., and Roturier, J. (2011). Qualitative Analysis of Post-Editing for High Quality Machine Translation. In *MT Summit XIII*, Xiamen, China.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 Joint Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Campbell, S. (2000). Critical Structures in the Evaluation of Translations from Arabic into English as a Second Language. *The Translator*, 6:37–58.

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. L. (2011). The process of post-editing: a pilot study. In *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation*, volume 41 of *Copenhagen Studies in Language*, pages 131–142, Fredriksberg. Samfundslitteratur.

Carl, M. and Kay, M. (2011). Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators. *Meta*, 56(4):89–111.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

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

Doherty, S. and O'Brien, S. (2009). Can MT Output be Evaluated through Eye Tracking? In *MT Summit XII*, pages 214–221, Ottawa, Canada.

Doherty, S., O'Brien, S., and Carl, M. (2010). Eye Tracking as an Automatic MT Evaluation Technique. *Machine Translation*, 24(1):1–13.

Edmundson, H. P. and Hayes, D. G. (1958). Research methodology for machine translation. *Machine Translation*, 5(1):8–15.

Elming, J., Balling, L. W., and Carl, M. (2014). Investigating User Behaviour in Post-Editing and Translation using the CASMACAT Workbench. In O'Brien et al. (2014), pages 147–169.

Englund Dimitrova, B. (2005). *Expertise and Explicitation in the Translation Process*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Federmann, C. (2012). Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Fiederer, R. and O'Brien, S. (2009). Quality and machine translation: A realistic objective? *Journal of Specialised Translation*, 11:52–74.

García, I. (2010). Is machine translation ready yet? *Target*, 22(1):7–21.

García, I. (2011). Translating by post-editing: Is it the way forward? *Machine Translation*, 25(3):217–237.

García, I. (2012). A brief history of postediting and of research on postediting. *New Directions in Translation Studies. Special Issue of Anglo Saxonica*, 3(3):292–310.

Gaspari, F., Almaghout, H., and Doherty, S. (2015). A survey of machine translation competences: insights for translation technology educators and practitioners. *Perspectives: Studies in Translatology*, 23(3):333–358.

Green, R. (1982). The MT errors which cause most trouble to posteditors. In Lawson, V., editor, *Practical experience of machine translation. Proceedings of a conference, London, 5-6 November 1981*, pages 101–104. North Holland Publishing Company, Amsterdam.

Guerberof Arenas, A. (2010). Project management and machine translation. *Multilingual*, 21(3):1–4.

Guerberof Arenas, A. (2014a). Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28:165–186.

Guerberof Arenas, A. (2014b). The role of professional experience in post-editing from a quality and productivity perspective. In O'Brien et al. (2014), pages 51–76.

Hu, C., Resnik, P., Kronrod, Y., Eidelman, V., Buzek, O., and Bederson, B. B. (2011). The value of monolingual crowdsourcing in a real-world translation scenario: simulation using Haitian Creole emergency SMS messages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT '11)*, pages 399–404.

Hutchins, W. J. (1986). *Machine translation: past, present, future.* Ellis Horwood, Chichester.

Hyönä, J. (1993). *Eye movements during reading and discourse processing*, volume 65 of *Psychological Research Reports.* University of Turku, Turku.

Jolliffe, I. (2002). *Principal Component Analysis.* Springer, New York, 2nd edition.

Jääskeläinen, R. (1999). *Tapping the process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating.* Joensuun yliopisto, Joensuu.

Koby, G. S. (2001). Editor's Introduction – Post-Editing of Machine Translation Output: Who, What, Why, and How (Much). In Krings (2001), pages 1–23.

Koby, G. S. and Champe, G. G. (2013). Welcome to the Real World: Professional-Level Translator Certification. *Translation & Interpreting*, 5(1):156–173.

Koehn, P. (2010). Enabling monolingual translators: Post-editing vs. options. In *NAACL HLT 2010: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings*, pages 537–545.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *7th Workshop on Statistical Machine Translation*, pages 181–190, Montréal, Canada.

Koponen, M. (2013). This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 1–9, Nice, France.

Koponen, M. (2016). Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort. *Journal of Specialised Translation*, 25. (in print).

Koponen, M., Aziz, W., Ramos, L., and Specia, L. (2012). Post-editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*, pages 11–20, San Diego, California.

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

Koskenniemi, K., Lindén, K., Carlson, L., Vainio, M., Arppe, A., Lennes, M., Westerlund, H., Hyvärinen, M., Bartis, I., Nuolijärvi, P., and Piehl, A. (2012). *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

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

Lacruz, I., Denkowski, M., and 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)*, pages 73–84.

Lacruz, I. and Shreve, G. M. (2014). Pauses and Cognitive Effort in Post-editing. In O'Brien et al. (2014), pages 246–272.

Lacruz, I., Shreve, G. M., and Angelone, E. (2012). Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 21–30.

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

Löffler-Laurian, A.-M. (1986). Post-édition rapide et post-édition conventionelle: deux modalités d'un activité spécifique. *Multilingua*, 5(2):81–88. and 5(4):225–229.

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

Mitchell, L., O'Brien, S., and Roturier, J. (2014). Quality evaluation in community post-editing. *Machine Translation*, 28(3):237–262.

Mitchell, L., Roturier, J., and O'Brien, S. (2013). Community-based post-editing of machine-translated content: monolingual vs. bilingual. In *Workshop Proceedings: Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 35–44.

Moran, J., Lewis, D., and Saam, C. (2014). Analysis of Post-editing Data: A Productivity Field Test Using an Instrumented CAT Tool. In O'Brien et al. (2014), pages 126–146.

Mossop, B. (2007). *Editing and Revising for Translators.* St. Jerome Publishing, Manchester and Kinderhook, 2nd edition.

Nieminen, T. (2013). Konekäännös kääntäjän apuna [Machine translation as the translator's aid]. Invited talk at Konekääntämisen teemailta [Machine translation theme night], organized by The Finnish Association of Translators and Interpreters SKTL, October 15, Helsinki, Finland.

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

O'Brien, S. (2006a). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14(3):185–205.

O'Brien, S. (2006b). Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. *Across Languages and Cultures*, 7:1–21.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, 17:55–77.

O'Brien, S., Balling, L. W., Carl, M., Simard, M., and Specia, L. (2014). *Post-editing of Machine Translation: Processes and Application.* Cambridge Scholars Publishing, Newcastle upon Tyne.

O'Curran, E. (2014). Machine Translation and Post-Editing for User Generated Content: An LSP Perspective. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas, Vol. 2: MT Users Track*, pages 50–54.

Padró, L., Collado, M., Reese, S., Lloberes, M., and Castellón, I. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *7th International Conference on Language Resources and Evaluation*, pages 3485–3490.

Paulsen Christensen, T. and Schjoldager, A. (2010). Translation-memory (TM) research: what do we know and how do we know it? *Hermes: Journal of Language and Communication Studies*, 44:89–101.

Plitt, M. and Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Popović, M., Lommel, A., Burchardt, A., Avramidis, E., and 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*, pages 191–198.

Rautio, J. and Koponen, M. (2013). MOLTO evaluation and assessment report. Technical report, MOLTO Project.

Rayner, K., Pollatsek, A., Ashby, J., and Clifton, C. (2012). *Psychology of Reading*. Psychology Press, New York and London, 2nd edition.

Robert, A.-M. (2013). Vous avez dit post-éditrice ? Quelques éléments d'un parcours personnel. *The Journal of Specialised Translation*, 19:29–40.

Rosti, J. and Jaatinen, H. (2015). Suomi-englanti-suomi konekäännöksen post-editointi [Finnish-English-Finnish machine translation post-editing]. Invited talk at the Meeting of the Machine Translation Special Interest Group, Kites Association, June 5, Helsinki, Finland.

Schwartz, L. (2014). Monolingual Post-Editing by a Domain Expert is Highly Effective for Translation Triage. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*, pages 34–44.

Schwartz, L. O., Anderson, T., Gwinnup, J., and Young, K. M. (2014). Machine Translation and Monolingual Postediting: The AFRL WMT-14 System. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194.

Silva, R. (2014). Integrating Post-editing MT in a Professional Translation Workflow. In O'Brien et al. (2014), pages 24–50.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2010). TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 22(2-3):117–127.

Sousa, S. C. M., Aziz, W., and Specia, L. (2011). Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Recent Advances in Natural Language Processing Conference*, Hissar, Bulgaria.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven, Belgium.

Specia, L., Cancedda, N., and Dymetman, M. (2010a). A dataset for assessing machine translation evaluation metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Specia, L. and Farzindar, A. (2010). Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry*, Denver, Colorado.

Specia, L., Raj, D., and Turchi, M. (2010b). Machine translation evaluation versus quality estimation. *Machine Translation*, 24:39–50.

Tatsumi, M. (2009). Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. In *MT Summit XII*, pages 332–333.

Tatsumi, M., Aikawa, T., Yamamoto, K., and Isahara, H. (2012). How Good Is Crowd Post-Editing? Its Potential and Limitations. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 21–30.

Tatsumi, M. and Roturier, J. (2010). Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship? In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC10)*, pages 43–51.

TAUS (2010a). Machine Translation Post-editing Guidelines. Technical report.

TAUS (2010b). Postediting in Practice. A TAUS Report. Technical report, TAUS BV.

Teixeira, C. (2011). Knowledge of provenance and its effects on translation performance in an integrated TM/MT environment. In *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation*, volume 41 of *Copenhagen Studies in Language*, pages 107–118, Fredriksberg. Samfundslitteratur.

Teixeira, C. S. (2014a). The Handling of Translation Metadata in Translation Tools. In O'Brien et al. (2014), pages 109–125.

Teixeira, C. S. C. (2014b). Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*, pages 45–59.

Temnikova, I. (2010). A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

Temnikova, I. and Orasan, C. (2009). Post-editing Experiments with MT for a Controlled Language. In *International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages*, Besançon, France.

Thicke, L. (2013). The industrial process for quality machine translation. *The Journal of Specialised Translation*, 19:8–18.

Underwood, N. L. and Jongejan, B. (2001). Translatability Checker: A Tool to Help Decide Whether to Use MT. In *Proceedings of MT Summit VIII*, pages 363–368.

van der Meer, J. and Ruopp, A. (2014). Machine Translation Market Report. Technical report, TAUS BV.

Čulo, O., Gutermuth, S., Hansen-Schirra, S., and Nitzke, J. (2014). The Influence of Post-Editing on Translation Strategies. In O'Brien et al. (2014), pages 200–218.

Vieira, L. N. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation*, 28(3):187–216.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Machine Translation Output. In *5th International Conference on Language Resources and Evaluation*, pages 697–702.

Zhechev, V. (2014). Analysing the Post-Editing of Machine Translation at Autodesk. In O'Brien et al. (2014), pages 2–13.

Appendix A
Author's contributions and division of labor

This appendix describes my own contributions as well as the division of labor during the collaborative work carried out as part of this dissertation project. The descriptions of co-authored articles have been approved by the respective co-authors.

For Koponen (2012), I was the sole author and conducted the work reported therein independently during spring 2012. The MT, PE and manual evaluation data used in the experiment was distributed to participants by the organizers of the WMT12 Workshop (Callison-Burch et al. 2012). My contributions involved determining the edit distance and effort evaluation parameters for selecting the potentially interesting sentences to be analyzed, as well as carrying out the manual analysis of the edits in these sample sentences.

Article Koponen et al. (2012) reports the first collaborative project, carried out in 2012 together with Dr. Lucia Specia (University of Sheffield), Wilker Aziz (University of Wolverhampton), and Luciana Ramos. The post-editing data used was collected by Lucia Specia and Luciana Ramos, with the main contribution of Luciana Ramos involving recruitment of the participants for the post-editing task and collection of the post-editing data. The data was collected using the PET tool created by Wilker Aziz (Aziz et al. 2012). Initial analysis and processing of the data for later analyses (error analysis and comparison of post-editors) was carried out mostly by Aziz and Specia. My main contribution involved the manual analysis: I selected the sample sentences based on the PE process data, planned the manual analysis method and carried out the analysis. The numerical analysis of post-editor variability was mainly carried out by Aziz. The results and conclusions from both analyses were discussed together by Specia, Aziz and myself, and I also participated in the description of the post-editor profiles observed. We all worked on the co-authored article, but mostly the writing was divided so that the sections *Introduction* and *Related work* were mainly written by Specia, the *Method* section was written in part by Aziz (description of the dataset and sub-section *Human variability in post-editing*) and myself (sub-section *Cognitive effort in post-editing*). The *Results* section was similarly divided between Aziz (*Human variability in post-editing*) and myself (*Cognitive effort in post-editing*, as well as some parts of the other subsection). The tables and figures in the article were produced by Aziz.

The collaboration with Lucia Specia and Wilker Aziz was continued during 2013 in the experiment described in Aziz et al. (2014). The process data used was received from the CASMACAT project (Elming et al. 2014). The original data was further processed mainly by Aziz, who also carried out the Principal Component Analysis described the article. My main contribution was selection of potential source text features to be investigated in the analysis as well as participation in interpreting the PCA results against the features used in this analysis. We all worked on the co-authored article, but most of the writing was divided so that the sections *Introduction* and *Related work* were mostly written by Specia, and I made additions related to translatability features. In the section *Sub-sentence level analysis*, the description of the data was mainly written by Specia. I wrote most of the section *Motivation* as well as the first part of the section *Source text and post-editing effort features* dealing with source text translatability features in general. The latter part of this section dealing with the generation of the specific source text features as well as the section *Principal Component Analysis* and the parts of the *Results* section describing observations on the data plots were written by Aziz. In the section *Results*, I made additions related to the interpretation of the translatability features. The plots in the article were produced by Aziz. The *Conclusions* section was jointly written by all authors.

For the article Koponen (2013) I was again the sole author and carried out the work independently in the summer of 2013. The MT and PE data used in the experiment were obtained through the EU MOLTO project, where I had participated in planning and organizing the evaluation procedures (Rautio and Koponen 2013). My contribution in this article was the analysis of the selections and edits made by the evaluators participating in one branch of the evaluation campaign.

The collaborative project reported in Koponen and Salmi (2015) was carried out together with Dr. Leena Salmi (University of Turku) in various stages between 2011 and 2014. The data used in this experiment was originally collected in a university course taught by Salmi at the University of Turku in the spring of 2011. My contributions included selection and preparation of the material to be post-edited and the instructions for the participants, planning the manual evaluation and error analysis methods used in the experiment, carrying out half of the manual evaluation error analysis, and most of the numerical analysis of the evaluation results. The contributions by Salmi included practical coordination of the data collection, selection of the sample sentences for the detailed error analysis, carrying out half of the manual evaluation and error analysis, as well as participation in the analysis of the evaluation results. We both worked on the co-authored article, but the writing was mainly divided so that I wrote most of the sections *Introduction*, *Discussion* and *Conclusions*, with some additions by Salmi. In the *Method* section, *Study setup* was written by Salmi, I wrote most of the sub-section *Evaluation of correctness*, while the section *Error analysis of simple and difficult cases* was written jointly.

For the article Koponen (2016) I was the sole author. The survey of literature was on-going from the start of the dissertation project.

Appendix B
Original Articles

The previously published articles forming part of this dissertation are reproduced at the end of the print version by permission of the copyright holders.