

Post-Editing Machine Translated Text in A Commercial Setting: Observation and Statistical Analysis

Midori Tatsumi, MSc.

Thesis submitted for the degree of Doctor of Philosophy

School of Applied Language and Intercultural Studies

Dublin City University

Supervisors:

Dr. Sharon O'Brien / Dr. Minako O'Hagan / Dr. Fred Hollowood

October 2010

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____
(Candidate)

ID No.: _____

Date: _____

ACKNOWLEDGEMENT

I would like to express my gratitude for the Enterprise Ireland Innovation Partnerships Programme in conjunction with Symantec Ltd. for giving me a rare opportunity to conduct research in both academic and industrial contexts while fulfilling my personal interest.

I would like to thank my academic supervisors, Dr. Sharon O'Brien and Dr. Minako O'Hagan, for guiding and enlightening me on how to perform research of academic value. This was an eye-opening experience in every sense for me, a person who had never been a resident of academia! I would also like to thank my industrial supervisor, Dr. Fred Hollowood, for maintaining the practical value of my research by questioning the meaning of each finding from an industrial view point.

I am grateful to Dr. Johann Roturier for providing constant and thorough technical help and astonishing expertise, and also to everyone I worked with at Symantec for helping me on many occasions and for friendship.

I would like to thank my family and friends in Japan for the continuous support during my long journey abroad.

To all friends I met in Ireland and especially in the post-grad rooms, thank you for your company!

Special thanks to Marian, for all the running sessions and half-marathons we did together. Running is one of the best things I have done during my study in Ireland.

And finally, thank you Frank, for constantly taking me outdoors to enjoy nature and giving me a 'Peaceful Easy Feeling'.

ABSTRACT

Post-Editing Machine Translated Text in A Commercial Setting: Observation and Statistical Analysis

Machine translation systems, when they are used in a commercial context for publishing purposes, are usually used in combination with human post-editing. Thus understanding human post-editing behaviour is crucial in order to maximise the benefit of machine translation systems. Though there have been a number of studies carried out on human post-editing to date, there is a lack of large-scale studies on post-editing in industrial contexts which focus on the activity in real-life settings. This study observes professional Japanese post-editors' work and examines the effect of the amount of editing made during post-editing, source text characteristics, and post-editing behaviour, on the amount of post-editing effort. A mixed method approach was employed to both quantitatively and qualitatively analyse the data and gain detailed insights into the post-editing activity from various view points. The results indicate that a number of factors, such as sentence structure, document component types, use of product specific terms, and post-editing patterns and behaviour, have effect on the amount of post-editing effort in an intertwined manner. The findings will contribute to a better utilisation of machine translation systems in the industry as well as the development of the skills and strategies of post-editors.

TABLE OF CONTENTS

PART I RESEARCH BACKGROUND

Chapter 1	Introduction	2
1.1	Machine Translation (MT)	2
1.2	MT and Post-Editing (PE)	3
1.3	IT Document As A Text Genre	5
1.4	Research Context	6
1.5	Structure of the Thesis	8
Chapter 2	Review of Post-Editing Studies	10
2.1	Definition of Post-Editing (PE)	10
2.2	Research on Strategies and Methodologies of PE	11
2.2.1	Human PE	11
2.2.2	Aided and automated PE	12
2.2.3	Statistical machine translation based post-editing (SPE)	14
2.3	Methods of Measuring Human PE Effort	15
2.3.1	Cognitive PE effort	16
2.3.2	Temporal PE effort	17
2.3.3	Technical PE effort	17
2.4	Findings from PE Effort Studies and Research Gaps	18
2.4.1	MT quality and PE effort	18
2.4.2	ST characteristics and PE effort	20
2.4.3	PE typology	22
2.4.4	Post-editor variance	24
2.4.5	Research settings	25
2.4.6	Analysis method	26
2.5	MT and TM	26
2.5.1	Combining MT and TM	26
2.6	MT and PE research in Japan	27
2.7	Formulation of Research Questions and Research Plans	30
2.8	Concluding Remarks	32

PART II CONDUCTING THE STUDY

Chapter 3	Research Methodology	34
3.1	Research Framework.....	34
3.2	Mixed Methods Research Design	35
3.2.1	‘What is happening?’	37
3.2.2	‘Is there a systematic effect?’	37
3.2.3	‘How is it happening?’	38
3.3	Observational Experiment (Data Collection Phase)	38
3.4	Operationalisation	39
3.4.1	Amount of PE effort (dependent variable).....	39
3.4.2	Amount of editing (independent variable)	41
3.4.3	Characteristics of source text (independent variable)	43
3.4.4	Types of PE operation (independent variable).....	44
3.5	Statistical Data Analysis	45
3.6	MT/PE Quality Assessment	46
3.7	Concluding Remarks.....	46
Chapter 4	Preliminary Study	48
4.1	Objectives.....	48
4.2	Methods.....	48
4.2.1	Preparation of test corpus.....	48
4.2.2	Post-editing session.....	49
4.3	Initial findings	50
4.3.1	Correlation between the amount of editing and PE effort.....	50
4.4	Effect of combined variables on PE effort.....	53
4.4.1	ST structure	53
4.4.2	ST length	55
4.4.3	Observed common PE operations	56
4.4.4	Dependency.....	58
4.4.5	Multiple linear regression analysis.....	59
4.5	Lessons learnt.....	63
4.6	Concluding Remarks.....	64

Chapter 5	Methods of Data Collection and Analysis	65
5.1	Measurement Technique	65
5.1.1	PE speed	66
5.1.2	Automatic Evaluation Metrics (AEM).....	68
5.1.3	Characteristics of source text	69
5.1.4	Types of PE operations	70
5.1.5	Observation of PE operation	70
5.1.6	Post-editors' attributes	71
5.2	Test Corpus Design	71
5.2.1	Sampling considerations	71
5.2.2	Procedures	74
5.2.3	Machine translation using Systran	77
5.3	Participant Post-Editors.....	78
5.3.1	Population	78
5.3.2	Sampling method	78
5.4	Experiment	79
5.4.1	Time and duration	79
5.4.2	Environment and equipment	80
5.4.3	Project kit	81
5.4.4	Instructions.....	81
5.4.5	Procedures	83
5.4.6	Research ethics.....	84
5.5	Questionnaire	87
5.5.1	Aim of questionnaire.....	87
5.5.2	Subjects and categories of questions.....	87
5.5.3	Designing questions	89
5.6	Concluding Remarks.....	91

PART III RESEARCH FINDINGS AND DISCUSSIONS

Chapter 6	Taxonomy Development for PE Operation Types.....	93
6.1	Method of Qualitative Analysis	93
6.2	Taxonomy Development for PE Operation Types.....	94

6.2.1	Coding	94
6.2.2	Categorising	112
6.3	Concluding Remarks	113
Chapter 7	Quantitative Analysis	114
7.1	Profile of Analysed Data	114
7.2	Summary statistics of PE speed and GTM	115
7.3	Correlation between the amount of editing and the PE speed	117
7.4	Effect of ST Characteristics	119
7.4.1	ST length	119
7.4.2	ST structure	121
7.4.3	Component parts of the document	123
7.4.4	Number of UI terms	125
7.4.5	Number of Technical terms	127
7.4.6	Complexity index by MT system	128
7.4.7	Conformity to controlled language rules	129
7.5	Effect of Post-Editing Operation Types	131
7.5.1	Edit frequency of each PE type	131
7.5.2	Number of visits	134
7.6	Effect of Combined Variables - Multiple Regression Analysis	137
7.6.1	Primary Independent Variable category	139
7.6.2	Source Text Characteristics category	140
7.6.3	PE Operation category	142
7.6.4	Post-Editor Variance category	144
7.6.5	Discussion	144
7.7	Comparison with TM match segments	145
7.7.1	TM match ratio that corresponds to MT output	145
7.7.2	Comparison of multiple regression analysis results	148
7.7.3	Edit frequency	150
7.8	Summary of Quantitative Analysis	155
7.9	Concluding Remarks	157
Chapter 8	Revisits, Outliers, and Post-Editor Variance	158
8.1	Revisits	158

8.1.1	Overview of revisits	159
8.1.2	General characteristics of PE during revisits	159
8.1.3	Between-post-editor difference in revising behaviour	160
8.1.4	Types and temporal effort of revisits	162
8.1.5	Section summary	163
8.2	Outliers in PE speed	163
8.2.1	Identifying the outliers to be further investigated	164
8.2.2	Comparison of profiles between outlier and normal cases	165
8.2.3	Qualitative investigation using screen capture video	166
8.2.4	Section summary	169
8.3	General Post-Editor difference	169
8.3.1	Non-edited segments	171
8.3.2	Observation of BB Flashback	173
8.3.3	Background data from questionnaire	177
8.3.4	Section summary	180
8.4	Overview of questionnaire results	180
8.4.1	PE training	180
8.4.2	PE methods	181
8.4.3	Reason for inefficient PE	182
8.4.4	MT and PE in general	182
8.4.5	Section summary	184
8.5	Concluding remarks	185
Chapter 9	Conclusions	186
9.1	Aims of the Study	186
9.2	Findings of the Study	187
9.3	Contribution to Knowledge	191
9.4	Limitations of the Study	192
9.4.1	MT system	192
9.4.2	Language pair	193
9.4.3	Methodology	193
9.5	Future Research	194

References	196
------------------	-----

APPENDICES

Appendix A	Project Manual	209
Appendix B	Part of Speech List	233
Appendix C	Post-Estimation Tests.....	236
Appendix D	Questionnaire	240
Appendix E	DCU Research Ethics Committee documents.....	251
Appendix F	Research Consent Form	254

LIST OF FIGURES

Figure 3.1 Overall research design combining two mixed methods designs	36
Figure 3.2 Concept of PE effort	40
Figure 4.1 Trados tools	49
Figure 4.2 Correlation between AEM scores and logarithm of PE speed (Post-Editor A)	51
Figure 4.3 Correlation between AEM scores and logarithm of PE speed (Post-Editor B)	52
Figure 4.4 Correlation between AEM scores and logarithm of PE speed (Post-Editor C)	52
Figure 4.5 Average PE speed by ST length and post-editors	56
Figure 4.6 Graphical representation of quadratic function	60
Figure 7.1 Correlation between average PE speed and GTM scores by post-editors	117
Figure 7.2 Correlation between PE speed and GTM scores for each sentence by post-editors	118
Figure 7.3 Average PE speed by ST length categories	120
Figure 7.4 Average PE speed by ST structures	122
Figure 7.5 Average PE speed by document component parts	124
Figure 7.6 Average PE speed by the number of UI terms	126
Figure 7.7 Average PE speed by the number of technical terms	128
Figure 7.8 Average PE speed by complexity index scores	129
Figure 7.9 Average PE speed by CL rule conformity	130
Figure 7.10 The number of sentences visited different times	134
Figure 7.11 Comparison between the average number of visits and PE speed	135
Figure 7.12 PE speed by number of visits	136
Figure 8.1 Correlation between PE speed and GTM scores (a sample extracted from Figure 7.2)	166
Figure 8.2 Average PE speed and GTM scores by post-editors	170
Figure 8.3 Average PE speed for GTM=1 segments and all segments	172
Figure 8.4 Number of GTM=1 cases and average PE speed	173
Figure 8.5 Example of screen layout	174
Figure 8.6 Experience in software manual translation and PE speed / GTM scores	177
Figure 8.7 Experience in PE and PE speed / GTM scores	178
Figure 8.8 Experience in TagEditor and PE speed / GTM scores	179
Figure C.1 Heteroscedasticity test	237
Figure C.2 P-P plot (left) and Q-Q plot (right)	238

LIST OF TABLES

Table 4.1	Summary statistics for PE speed (words/min)	50
Table 4.2	Summary statistics for automatic metric scores (N=1,425).....	51
Table 4.3	Difference in correlation levels between sentence structures	55
Table 4.4	Average PE speed (words/min) for different ST length groups.....	55
Table 4.5	Regression analysis results of PE speed by ST structures	61
Table 5.1	Composition of TM matches	76
Table 6.1	An example of aligned information extracted from TER and MeCab.....	96
Table 6.2	Taxonomy of PE operation types.....	112
Table 7.1	Profile of analysed text	115
Table 7.2	Average PE speed and GTM scores.....	115
Table 7.3	Sentence length by sentence structure	121
Table 7.4	Mean GTM scores and PE speed by sentence types	122
Table 7.5	Frequency of each document component part in the test corpus	123
Table 7.6	Cross tabulation of document component parts and ST structures	124
Table 7.7	Frequency of PE operation by post-editors.....	131
Table 7.8	Example of PE operation classification	133
Table 7.9	Average number of visits and PE speed.....	135
Table 7.10	Regression analysis results of the effect on PE speed	139
Table 7.11	Comparison of average editing speed between MT output and TM matches.....	146
Table 7.12	Comparison of average GTM scores between MT output and TM matches	147
Table 7.13	Regression analysis results of the effect on editing speed (MT–TM comparison)	149
Table 7.14	Average edit frequencies by part of speech and technical effort (MT).....	151
Table 7.15	Average edit frequencies by part of speech and technical effort (TM).....	152
Table 7.16	Edit frequencies by operation types	154
Table 8.1	Number and types of revisits by post-editors.....	159
Table 8.2	Revisits timing by post-editors	161
Table 8.3	Types and temporal effort of revisits	162
Table 8.4	Number of normal and outlier cases (segments).....	165
Table 8.5	Profiles of fast, normal, and slow PE cases	165
Table 8.6	Number of GTM=1 items and average PE speed	171

LIST OF ABBREVIATIONS

AEM	Automatic Evaluation Metrics
AMTA	Association for Machine Translation in the Americas
BLEU	BiLingual Evaluation Understudy
CAT	Computer Aided/Assisted Translation
CBDF	Chi divided By Degrees of Freedom
CL	Controlled Language
CNA	Choice Network Analysis
CNGL	Centre for Next Generation Localisation
EC	European Commission
FIGS	French, Italian, German, and Spanish
GIM	Global Information Management
GTM	General Text Matcher
IAMT	International Association for Machine Translation
IT	Information Technology
IQR	InterQuartile Range
JEIDA	Japan Electronic Industry Development Association
LISA	Localization Industry Standards Association
MLR	Multiple Linear Regression
MT	Machine Translation
NTI	Negative Translatability Indicators
PAHO	Pan American Health Organisation
PE	Post-Editing
PoS	Part of Speech
RBMT	Rule-Based Machine Translation
SD	Standard Deviation
SMT	Statistical Machine Translation
ST	Source Text
TAP	Think Aloud Protocol
TAUS	Translation Automation User Society
TDA	TAUS Data Association
TER	Translation Edit Rate
TM	Translation Memory
TT	Target Text
UD	User Dictionary
UI	User Interface
XML	eXtensible/eXtended Markup Language

PART I

RESEARCH BACKGROUND

Chapter 1 INTRODUCTION

1.1 Machine Translation (MT)

The use of machine translation systems has become more and more prevalent across various user categories from individual home users, who wish to understand the contents of foreign information quickly, to governments and large-scale multinational corporations that need to distribute a large amount of information in many languages in a cost-effective and timely manner.

Already in 2003, Hutchins reported that MT systems had been widely employed by large multinational organisations, including Ericsson, SAP, Corel, Ford, General Motors, Berlitz, Xerox, and so forth, over the previous couple of decades (Hutchins 2003). He stressed that one of the necessary conditions for effective MT implementation was that the organisation expected a large volume of translation, because the preparation and maintenance for MT deployment required a great amount of work, which might only be justified where a large amount of translation was expected to be processed by MT. He stated that large organisations that translated 100,000 pages or more annually could in some cases expect a cost reduction of 40-50% in addition to faster production, whereas smaller scale implementation of MT might only result in faster production, and not a very large cost saving. MT systems have since even increased in popularity partly because of the improvement of MT performance along with the need for more cost-effective translation. In 2008, SDL Research reported even more widely growing interest from global businesses in employing MT (SDL Research 2008).

Among large multinational organisations, MT has played an especially significant role for the IT industry. Txabarriaga et al. report that software publishing is the top industry in translation in Europe among 27 industry categories; nearly 10% of the entire translation market is for software publishing (Txabarriaga et al. 2009). The growing interest can be observed not only by the figures shown in the survey, but also by various movements. In the MT Summit in 2009, one of the world's largest conferences on MT, a number of studies were reported by research groups in IT companies, including Microsoft, IBM, Adobe, and Symantec. Also, a number of organisations, such as LISA

(Localization Industry Standards Association), TAUS (Translation Automation User Society), and CNGL (Centre for Next Generation Localisation) have recently been conducting various research projects on translating and localising technical documentation, for which one of the main providers is the IT industry. In addition to the need for speed and cost-effectiveness, there is another reason why MT is especially popular in the IT industry, that is, the amenability of the text to MT. The user manuals and help contents of IT products tend to contain many repetitive or similar sentences with limited variety of structures, which is considered to be suitable to be translated by MT (DePalma & Kelly 2009).

1.2 MT and Post-Editing (PE)

MT systems translate text in a speedy manner, but they do not always produce satisfactory results. Flanagan stated that MT output might fit the gisting purposes for perishable information, and she reported that 85% of CompuServe's MT output was published without post-editing, and specifically, only the non-edited version was available in their online discussion forum (Flanagan 1997). But this was a rare example; in most cases where publishing quality was required, MT output needed human intervention, 'post-editing (PE)', to raise quality to an acceptable level (Allen 2001, Hutchins 2003, Schäfer 2003). And this holds true to this day. Despite the efforts by MT developers and users to raise the quality of MT output, by populating user dictionaries (UD) of RBMT systems with company-specific and industry-specific terms, employing controlled language (CL) rules and automated pre-editing techniques to author MT-friendly source text (ST), and developing various automatic or semi-automatic post-processing techniques to implement corrections for repetitive errors or fully automate PE, MT output today still needs to be post-edited by humans in order to produce publishing quality translation (Roturier 2009, TAUS 2010).

DePalma & Kelly (2009) state that even when the MT output needs human PE, it is generally faster and cheaper than human translation, and when the cost is the same, MT plus PE achieves faster turnaround. Also, some recent studies have shown that the quality of the final product of MT plus PE can in some cases exceed the quality of human translation (Fiederer & O'Brien 2009, Koehn 2009), which may further justify

the increasing employment of this workflow. Nonetheless, it cannot be denied that human PE “increases the bill” (DePalma & Kelly 2009: p.8) and is one of the biggest issues in cost-effective and time-saving use of MT (Itagaki et al. 1999). In order to overcome this issue, the PE process needs to be further optimised (TAUS 2010). This calls for continuous effort for extensive research into PE.

As will be discussed in detail in Chapter 2, PE has been broadly researched since the 1980s especially in Europe. Suggestions have been made regarding PE methods, PE environments, and automated PE. However, many of these studies are anecdotal and lack objective measurement of effectiveness. There have also been studies in which the impacts of various factors and methods were tested against the amount of PE effort. However, such studies have often been conducted as controlled experiments, and thus did not observe PE effort in a commercial work environment. There is a need for a large-scale study in a real-life environment with statistical analysis of results. In addition, there has been little research on post-editing of English to Japanese translation despite the fact that the Asian languages, and especially Japanese, have been strategically important target languages for MT (DePalma & Kelly 2009, Japan Translation Federation 2009) and that there have been a number of governmental and private MT research and development initiatives in Japan.¹

The overall purpose of the present study is to answer the fundamental question ‘What determines the amount of PE effort?’, which might appear to be obvious: is it not just the amount of editing, which is ultimately determined by the quality of MT output? However, prior studies have shown that this is not that straightforward an issue; the relationship between them are not linear (see section 2.4.1 for a detailed discussion), thus we need to conduct detailed analyses from various viewpoints. The present study also aims to fill the above mentioned research gaps by reproducing the commercial work environment of professionals with popular software programs and commercial texts translated from English into Japanese, and analysing the results using advanced statistical methods.

¹ MT development, education, and publicity effort by Asia-Pacific Association for Machine Translation, various research projects by Language Translation Group of National Institute of Information and Communication Technology, MT dictionary development support and MT quality improvement on patent translation by Japan Patent Information Organisation, a community-based Web MT site by Oki Electric Industry Co., Ltd, to name a few.

1.3 IT Document As A Text Genre

The subject domain of this research falls into the genre ‘technical documents’ and its subgenre ‘IT documentation’, or more specifically, ‘software user guides’, if we assume these are recognised genres. This is a relatively new text category. Biber (1988), in using the LOB Corpus (Lancaster-Oslo-Bergen Corpus of British English) (Johansson et al. 1978) for analysing written English, discussed fifteen genres, namely, Press Reportage, Press Editorials, Press Reviews, Religion, Skills and Hobbies, Popular Lore, Biographies, Official Documents, Academic Prose, General Fiction, Mystery Fiction, Science Fiction, Adventure Fiction, Romantic Fiction, and Humour, none of which could be considered to include technical documents or IT documentation. This is understandable since these documents were not read by a wide audience at the time when the texts included in the LOB Corpus were published (1961). However, since then, as personal computers have become widespread, technical documents, especially IT-related texts, have become an important text genre as they are increasingly read and used by a wide variety of readers.

While Markel (2003) defines technical documents as the text whose content addresses specific users and helps readers solve problems, Byrne sees software documentation as one of the subcategories of technical documents, and further defines them as texts that explain the concepts, procedures, and other related information of “non-hardware components of a computer” and address people with different levels of knowledge and skills in the relevant field (Byrne 2006: p.53). He specifically focuses on software user guides, and explains their function, audience, desirable quality, and structural, linguistic, and visual characteristics in considerable detail (ibid: pp. 57-96). This may be an indicator of the recent establishment of this text genre as an important element in translation studies.

Software documentation, naturally, is regarded as an important text genre especially from an industrial point of view. TAUS recently founded an association called TDA (TAUS Data Association) for the purpose of industry-wide language data sharing, in which a number of IT-related companies participate. Its pilot project conducted in 2009 involved over fourteen million words of translation memory (English - French) from five computer software companies and investigated how data sharing between different

companies could streamline the translation process of software strings and documentation. The report from the pilot project suggests that 16% of the ST of a new translation project of a company can be translated by TM matches within the 80-100% match category created by other companies in the same industry (TAUS 2009). This demonstrates that a corpus-based study in IT/software text genre is becoming increasingly important.

1.4 Research Context

The present study is funded by a joint Enterprise Ireland²-Symantec³ Innovation Partnerships Fund (IP/2006/0368/E). Enterprise Ireland is a government agency responsible for supporting Irish businesses. The Innovation Partnerships programme aims at supporting collaborative research between Irish higher education institutes and companies, so that the companies can exploit research expertise and resources to develop innovative products, processes, and services, while researchers who participate in the programme are given opportunities to pursue their research interests while balancing academic excellence and real-life technical challenges directly pertinent to the industry.

This programme was especially ideal for the researcher of the present study, who has worked in the localisation departments of WordPerfect and Novell, both US-based IT companies, since 1991, and as an IT-specialised freelance translator since 1996. During her career, she was directly exposed to the advancement of translation technologies, including various computer aided translation (CAT) tools, such as online glossary and translation memory (TM), and MT, and developed an interest in streamlining large scale translation workflow by means of translation technologies as well as understanding human processes, especially the roles of translators, involved in the workflow. The researcher was given an opportunity by this funded programme to address the topic relevant to the industry from both the academic and industrial points of view, building on academic research methodologies while working closely with Symantec's R&D team in their office, supported by their financial aid and human and electronic resources.

² Enterprise Ireland: www.enterprise-ireland.com [Last accessed: 19/10/2010]

³ Symantec: www.symantec.com [Last accessed: 19/10/2010]

Symantec is one of the world's largest multinational software companies with headquarters in California and offices in more than 40 countries. Symantec develops and sells security, storage, and systems management solutions, and their customers range from consumers and small businesses to global organisations. Symantec produces more than 20 million words of documentation annually, portions of which are translated into roughly 30 languages. In order to produce such a large amount of translation in a cost-effective and timely manner, Symantec introduced an MT system, Systran, in 2004 to combine with the already employed TM system.

Systran is a rule-based machine translation (RBMT) system,⁴ which has been one of the cutting edge MT systems (especially in Europe) with most major European languages and some other languages, including Asian languages, covered. Systran is employed by a number of global enterprises, including Symantec, Cisco, and EADS,⁵ Internet portals, such as Yahoo!®, Lycos®, and AtlaVista™, and public agencies like the US Intelligence Community and the European Commission (EC). The reason that Symantec chose Systran as their corporate-wide standard MT system was primarily because, at the time of their planning for the introduction of MT in 2003, Systran was the only MT system to their knowledge that could handle all seven languages they needed, namely, French, German, Italian, Spanish, Brazilian, Simplified Chinese, and Japanese. The secondary reason was that Systran offers an extensive opportunity to customise the system to meet the specific needs of the company, including user dictionaries and style sheets.

The fundamental question of the present study 'What determines the amount of PE effort?' was inspired by the needs of Symantec. Since the introduction of an MT system, Symantec achieved a significant cost reduction in translation production.⁶ Still, PE

⁴ There are three major types of MT systems: Rule-Based MT, Statistical MT, and Example-Based MT. Rule-Based MT systems translate according to the defined grammar rules, and Statistical MT systems build statistical models based on the bilingual parallel corpora and apply the models when translating the text, while Example-Based MT systems use bilingual parallel corpora as the main knowledge base on the fly during the translation.

⁵ European Aeronautic Defence and Space Company N.V. (<http://www.eads.com> [Last accessed: 19/10/2010])

⁶ Brennan, S., in a keynote speech "Social Networking: Integrating the Customer into Content Creation and Localization" at LISA Forum, Dublin 2008

represents a significant cost as this process needs to be done by humans. More importantly, the cost-reduction has not occurred equally among all target languages. The highest reduction has been achieved for French and Spanish (60%), while German enjoys a reduction of between 40-50%, but Chinese and Japanese have seen only a 20-30% reduction. Chinese and Japanese are strategically important languages for Symantec, partly because of the size of the (potential) market, and also because of the fact that most products need to be translated into these two languages whereas, in European countries, some of the products can be sold in English. This calls for extensive research on streamlining MT + PE workflow in these language pairs.

Since the present study is firmly rooted in Symantec's context, it needed to be executed within a certain framework, including the choice of MT system and the text used in the study. As will be discussed in detail in Chapter 5, Systran, an RBMT system, is used solely throughout the present study; despite the fact that statistical machine translation (SMT) systems are also becoming increasingly popular especially in the field of MT research and among some commercial entities, including Google, Autodesk, and Microsoft, and that Symantec itself has also been researching the possibility of introducing SMT technology, Symantec still finds Systran the only practical option to satisfy their translation quality requirement. The text for translation was extracted from Symantec's documentation. The study was conducted by collaborating with not only Symantec's R&D team in Dublin, but also with the Symantec Japan office. The researcher has visited the Japan office twice during the course of the research, working with local staff, conducting the data collection in Japan, utilising their human resources, vendor contacts, workplaces, and expertise in Japanese PE.

1.5 Structure of the Thesis

The thesis is presented in three parts. *Part I: Research Background* includes two chapters. Chapter 1, which is this chapter, discussed the current situation related to this study, and addressed contextualisation. Chapter 2 reviews prior studies that are most relevant to this, which informed the research questions.

Part II: Conducting the Study comprises three chapters. Chapter 3 presents the theoretical research methodology employed in this study. Chapter 4 reports on the preliminary study, following which Chapter 5 explains practical methods for data collection and analysis, that have been refined based on the lessons learned from the preliminary study.

Part III: Research Findings and Discussions consists of three chapters, each of which deals with one of three phases of analysis. Chapter 6 presents the process and the results of the first of the two qualitative analysis phases: development of a PE taxonomy. Chapter 7 presents the quantitative analysis results, which is the main analysis phase of the present study. Chapter 8 discusses the results of the second of the two qualitative analysis phases, which serves as supplemental information to the quantitative results. After part III, Chapter 9 concludes the study, and points to future research opportunities.

Chapter 2 REVIEW OF POST-EDITING STUDIES

This chapter begins by defining post-editing (PE) in section 2.1, then moves onto a review of the past and ongoing PE research. First, section 2.2 reviews studies on strategies and methodologies of PE, then section 2.3 introduces different approaches to studying human PE effort that have been developed over the last two decades. Section 2.4 focuses on findings from the relevant studies and points out research gaps. Section 2.5 discusses the issues of the relationship between MT and TM, and section 2.6 reviews MT and PE research in Japan. Finally, research questions will be formulated in section 2.7.

2.1 Definition of Post-Editing (PE)

Post-editing (PE) is “by far most commonly associated as a task related to MT” (Allen 2003: p.297), and generally defined as the act of correcting and editing of the text translated by an MT system (Austermühl 2001, Allen 2003 etc.), or more concisely, “repairing texts” as the title of Krings’s canonical book aptly summarises (Krings 2001). However, the types and the extent of required correction and editing depend on a number of factors, including the intended audience, the volume and the time constraint of the project, the expectation regarding the quality, and so forth. Sometimes only accuracy is needed, but sometimes stylistic refinement is required (McElhaney & Vasconcellos 1988, Austermühl 2001, Allen 2003, TAUS 2010). The process of PE also has some variations, including editing tools, the type and the form of dictionaries or glossaries provided, whether or not PE is done monolingually on the translated text or bilingually consulting with the ST, the file format of the source and machine-translated text, etc. (There used to be the need to distinguish between paper-based and electronic document-based PE, but nowadays PE is conducted on the computer in most, if not all, cases.) In the present study, PE is defined as, using a specific ST and MT output pair, any act of editing performed on the MT output so that the final product accurately conveys the information in the ST and conforms to the grammar of the target language. When MT output already meets these conditions and does not need any editing, any act of confirming it, such as reading the ST and MT output, is also considered as PE.

2.2 Research on Strategies and Methodologies of PE

PE is not a new topic in MT-related research; it was studied quite eagerly already in the 1980s especially in Europe where international organisations had a need to share information rapidly in many languages. The initial motivation was to develop effective strategies and methodologies of human PE in order to make the most efficient use of MT output.

2.2.1 Human PE

The concept of ‘rapid PE’ was developed in the EC in the 1980s as a method to perform a minimum amount of corrections to the text in order to process a large amount of documents for gisting purposes. The Directorate-General for Translation (DGT) of the EC claims to be the world’s largest translation service organisation, whose annual production grew from about 30,000 pages in 1990 to about 1,500,000 pages in 2005, 860,000 of which were translated by an MT system (Directorate-General for Translation 2008). Since the 1980s, the EC recognised that the best use of MT was to provide quick translation to those who agreed to accept low quality translation (Wagner 1985). To produce not exactly human translation quality, but sufficiently accurate translation, they developed a way of rapidly post-editing the text. The focus of rapid PE was on speed, and post-editors were advised to make minimum changes to finish PE of one page of text within half an hour while maintaining “comprehensibility and reasonable accuracy” (Wagner *ibid*: p.203), though the types of PE operations that should have been carried out or avoided were not described in detail.

McElhaney & Vasconcellos (1988), in reporting the PE experience in the Pan American Health Organisation (PAHO), made some suggestions for efficient PE, such as:

- Work on the text from left to right
- Avoid major rearrangement of the sentence
- Make use of mechanical aid, such as mouse and ‘search’ function

They also reported that after a month of practice, the average throughput of post-editors was 6,000 words per day.

Strategies and methodologies of human PE had been developed mainly in order to reduce the cost and time, but those were not the only concerns about human PE. PE effort often inevitably involves repetitive and tedious corrections of small mistakes (Wagner 1985, Allen 2003 etc.), which is exhausting, but hardly rewarding for human editors. To address these issues, a variety of efforts for minimising PE tasks have been made.

2.2.2 Aided and automated PE

Some proposals have been developed on interactive PE environments. For example, Allen (2001) reviewed and tested a software program called Reverso, with which MT, UD building, and PE can all be done in an interactive manner. The ST is translated automatically, then a quick review by a human is performed to identify “unknown, non-translatable, and mistranslated terms” (ibid: p.27), then the text is retranslated accordingly. The human PE process is performed on the integrated on-screen environment, which was said to speed up the PE process.

One of the most recent attempts to develop an interactive and mixed translation and PE environment is Caitra (Koehn 2009), which offers three types of ‘assistance’. *Prediction*, after a user has started translating the ST by typing one or more words, predicts the next possible few words to offer a sentence completion function. *Translation options* provides phrase-by-phrase translation suggestions, based on an SMT system, from which a user can choose one or discard all. *Postediting* simply shows MT output, which a user can edit. Koehn (ibid) has conducted a user study on using this environment hiring ten non-professional translators and using French as the ST and English as the TT, and reports that on average, all these types of assistance helped to increase speed and the quality of the translation compared to non-assisted human translation. Among three types of assistance, the PE module was most effective in cutting the translation time, increasing the speed by 39% compared to non-assisted translation.

There have also been attempts to automate some of the PE tasks that do not necessarily require human intervention. Vasconcellos (1987) and Allen (2003) introduced how PE was automated at PAHO. In PAHO, although post-editors are not given specifically designed PE training, they are provided with some techniques to accelerate the PE

process. One example is global search & replace, which is used to replace certain words or sequences of words, sometimes with format characters within one document. Another example is a set of scripts called ‘macros’ to deal with repetitive actions, such as moving a chunk of text. As an effort to automate PE more extensively, Wagner (1985) reported that in the EC they were using a customised MT system sub-routine to automatically convert present tense into reported speech when translating minutes of EC meetings written in French and translated into English to accommodate the difference in stylistic conventions between those two languages.

Ideas to combine manual analysis and statistical analysis in order to automate some PE tasks have also been suggested. For example, Knight & Chander (1994) focused on an article (a/an/the) insertion task for Japanese to English MT. Since the Japanese language does not have articles, article selection could be problematic when translated to English.⁷ To build the automatic PE module, they first analysed the features of noun phrases of over 400,000 phrases from Wall Street Journal text, and characterised them into binary rules. For the cases where the article must be determined taking into account the feature of the context, they also employed a statistical method to train the program. The result they obtained from the initial testing was 78% accuracy.

Elimination of manual analysis was the next step. Allen & Hogan (2000) has suggested a prototype for an Automatic Post-Editing (APE) module, which automatically learns frequently occurring corrections on MT output from previously post-edited documents, and applies the same correction to newly machine-translated text. This differs from the case where humans manually analyse the post-edited text and construct the modification rules, since all the steps will be performed automatically. The changes will be applied to the texts before they are passed to human post-editors. This concept of using monolingual parallel corpora of pre-PE output and post-PE target text (TT), allowing a module to learn from differences automatically, and applying the changes to the new text, is now being realised by means of statistical PE techniques, which will be discussed in more detail in section 2.2.3.

⁷ This has also been tackled by MT developers, for example, (Bond et al. 1994).

There have been some attempts to develop fully-automated PE modules. Povlsen & Bech (2001), having realised based on a survey that word order-related errors irritate post-editors greatly, put their focus on this particular problem and developed an automatic PE module for English to Danish translation called ‘Ape’. Ape first detects the segments that have been incorrectly translated by MT, and then fixes word order. The word reordering is performed according to predefined rules. The results of quantitative and qualitative evaluations showed that PE performance improved by 10-15%, and the post-editors commented that the number of problematic sentences was reduced considerably.

There has also been research into software programs that automatically learn from PE data and embed the acquired knowledge into the MT system itself. For example, the system proposed by Llitjós et al. (2007) consists of an online PE tool, from which error correction data are obtained, and a rule refiner, which determines the source of errors among MT system rules, and fixes them automatically. This system is not to help the human PE process itself, but rather to use the PE knowledge to refine MT systems, though it of course, in turn, could help to reduce human PE effort. These ideas and studies led to SMT-based PE.

2.2.3 Statistical machine translation based post-editing (SPE)

Automation of PE by means of SMT techniques is one of the rapidly developing areas of MT-related research. While more ‘traditional’ RBMT systems translate text according to the grammatical rules defined in their algorithms, SMT systems learn the translation pattern from existing STs and TTs, and apply the learned rules to translate new text. Most of the SPE techniques work in a somewhat similar manner; they learn from a set of target language parallel corpora consisting of raw MT output text and either post-edited or human-translated text. There have been a number of reports of combining a rule-based MT system and a SPE module. Ehara (2007a) reports on their system which is specifically developed to work on English to Japanese patent translation. Based on the experiment and evaluation, he concluded that the rule-based part of the system is good at handling structural transfer, and the statistical part of the system is good at lexical transfer of technical terms. Simard et al. (2007a) and Simard et al. (2007b) have conducted experiments on combining RBMT and SPE between

English and French in both directions and obtained positive results in terms of automatic evaluation metric (AEM) scores, which compare the resulting TT against approved quality translation, and measure the difference between the two texts.

Dugast et al. (2007) have conducted a more extensive evaluation, which consists of qualitative and linguistic evaluation by categorising the changes made by SPE to see if the changes caused improvements or degradations. They have concluded that SPE produced significant improvements in terms of lexical choice of locally better received terms, some improvements in grammaticality, especially in relation to determiners, but no remarkable improvements in sentence restructuring and word or phrase reordering. Tatsumi & Sun (2008) also reported that for both Chinese and Japanese, the changes made by SPE are largely limited to the word level, and sentence level editing seemed difficult to achieve using SPE. Roturier (2009) points out that deploying SPE poses problems in controlling the improvement/degradation ratio and the effect is different depending on the language pairs.

2.3 Methods of Measuring Human PE Effort

While a number of strategies and methodologies have been suggested, detailed investigations into the nature of human PE tasks had been largely neglected, or confined to mostly subjective and anecdotal observations, thus a shift to systematic observation and objective evaluation was in need (Krings 2001, O'Brien 2006b). In filling this gap, Krings conducted an extensive study on human PE processes, emphasising the importance of distinguishing the *product* (text) and the *process* (path that has been taken to produce the text). He pointed out that the effort in PE processes cannot be discussed in a one-dimensional manner. He proposed three different aspects of PE effort: temporal, cognitive, and technical. Temporal effort is most visible and easily measured. It is the indicator of the amount of PE effort that highlights the effort intensiveness of correcting each different type of deficiency. Cognitive effort is the decisive variable that explains the temporal effort, but since it cannot be measured directly, one needs to devise a method to reveal post-editors' mental processes. Technical effort is the mechanical process of implementing the corrections once the post-editor knows what corrections to

make. It can be described as insertion, deletion, and reordering. Among the three PE aspects, he put a great emphasis on investigating ‘cognitive’ effort.

2.3.1 Cognitive PE effort

Krings employed Think Aloud Protocol (TAP) (2001) as a means of evaluating cognitive effort. This method requires post-editors to verbalise all thoughts they have during the course of PE, so that the reasons for actions and non-actions can be made known to the researcher. While it certainly enables the recording of a lot of information on cognitive effort of post-editors, it has disadvantages, including 1) the use of this method itself creates an unusual setting, which may distort the data, 2) the act of thinking aloud can slow down or alter the normal cognitive process, 3) there is no guarantee that the verbalised statement is the accurate representation of subject’s cognitive process (Jakobsen 1998). O’Brien (2005), on the other hand, analysed cognitive PE effort by combining two methods: Translog and Choice Network Analysis (CNA). Translog is a software program that records all key strokes and mouse movements, as well as time, which was developed by Jakobsen (ibid) based on the assumption that the translation behaviour can be quantitatively analysed by observing the technical process involved in it. CNA (Campbell 2000) is a method of determining translation difficulty based on the diversity of translations produced by a number of translators, on the grounds that the number of possible translations for a given ST indicates the complexity of cognitive process required to produce a translation. While Translog is a more objective and non-invasive way to help capture the cognitive effort compared to TAP, a disadvantage of this method is that it cannot explain the reasons for pauses, whether they are caused by hesitation, thinking, lack of keyboard skills, or something else, though it could be compensated by triangulating with CNA. In order to achieve more direct and detailed measurement of cognitive effort in translation, O’Brien (2006a) tested eye tracking. An eye tracker records a subject’s eye movements on screen and pupil dilation in order to measure the cognitive load of the subject. It is based on the assumption that when eyes are fixed on an object the brain is engaged in some cognitive processing. She found that percentage change in pupil dilation and processing speed had a strong correlation. An eye tracker was also later tested as a method of capturing readers’ mental process when reading machine translated text (O’Brien 2006a, Doherty & O’Brien 2009, Doherty et al. 2010), which could be used as

an MT output quality measurement. They found that this method has satisfactory level of correlation with human evaluation of MT output.

These metrics have helped or may help with measuring the human PE effort and understanding the task in more detail. However, cognitive effort is not unequivocally visible, and there is no unanimously agreed upon measurement for cognitive effort. A quicker, simpler, and more commonly used metric for overall analysis may be in demand for commercial users, who constantly deal with projects that require translation of hundreds of thousands of words. One possible metric that can meet this need is ‘time’.

2.3.2 Temporal PE effort

The importance of time in terms of PE effort has been recognised from an early time. “The time required for post-editing naturally influences the *cost* of machine translation” (Wagner 1987), which emphasises the importance of understanding temporal PE effort in the context of commercial settings. Krings (2001), while emphasising the significance of cognitive effort in PE process research, also pointed out that temporal PE effort is an economically significant aspect of PE effort. Time has been employed in a number of studies as a measurement of the amount of PE effort, usually in the form of speed (the number of words processed in a minute, hour, etc.) (Krings 2001, O’Brien 2006b, Guerberof 2008, Plitt & Masselot 2010).

2.3.3 Technical PE effort

Technical effort can be examined from different points of view. O’Brien (2006b) measured technical effort by a process oriented approach using Translog, which keeps track of all the keyboard and mouse operations performed by post-editors. Groves & Schmidtke (2009) took a more product oriented approach by analysing the difference between two texts: the one before PE and another after PE. They compared SMT output and human post-edited text, and traced the minimum path between the two. As a result of the study, they found out that the most common PE operations for English to German and French translations are insertions, deletions, and alterations of function words, such as determiners and punctuation. Their study has made a contribution in identifying PE patterns in a systematic way.

2.4 Findings from PE Effort Studies and Research Gaps

A number of studies on PE effort compare the amount of human effort between traditional translation and PE of MT output. Krings (2001) reported from his study (English into German, French into German, and German into English) that the relative increase in speed when using the MT + PE solution compared to human translation from scratch was only +7% when post-edited on paper, and -20% when post-edited on screen. As a result, he concluded that the MT + PE solution does not help to reduce the amount of cognitive or temporal effort compared to human translation. However, this is the result from his experiment in 1989-1990, since which time MT quality has improved greatly, and translators and post-editors have acquired skills in using computers.

More recent studies have shown mostly opposite results. O'Brien (2006b) reports that, in her experiment (English into German translation of authoring software documentation), the average speed of PE was 17.59 words/min, while that of translation was 13.63 words/min, which suggests that PE is nearly 30% faster than translation. In a study Guerberof (2008) conducted to compare the productivity and the quality between post-editing of MT output and editing of TM fuzzy matches (English into Spanish translation of supply chain software documentation), the average PE speed was 13.86 words/min, which showed a productivity gain of 25% compared to human translation. Plitt & Masselot (2010) report even more striking results from their study (English into FIGS (French, Italian, German, and Spanish) translation of design and engineering software documentation): that the productivity gain from translation to PE was 74%. These results suggest that the situation has changed drastically from the time of Krings, and it now is proven that at least in the language pairs mentioned above, MT + PE is faster than full human translation. However, we still need to understand what factors affect PE speed.

2.4.1 MT quality and PE effort

Krings (2001) examined the relationship between the quality of MT output and the subsequent PE effort. The quality is measured by a human evaluator by using a five-level rating scale from 'poor' to 'good', while the amount of PE effort was measured by time and TAP. He found that MT quality had a clear impact on PE effort, but rather in a

surprising way. MT output quality and the amount of editing performed during the PE process were more or less in a linear relationship; the better the MT output quality, the smaller the textual difference between MT output and post-edited TT. However, the relationship between MT output quality and the temporal and cognitive PE effort was non-linear. He observed that the frequency of ST reading activities was higher with medium quality MT output than not only with good quality MT output but also with poor quality MT output. He also noted that the TT production was more difficult with medium quality MT output than not only with good quality MT output but also with poor quality MT output. From these observations, he concluded that the “post-editing effort is not highest for poor machine translations, but rather for medium-quality machine translations.” (Krings 2001: p539). This, in turn, suggests that the textual difference may not directly reflect the amount of PE effort.

The results of an evaluation project Ramirez & Haller (2005) conducted partly support Krings’s observation. They used professional translators as human evaluators on the one hand, and AEM on the other, and compared the results. Human evaluators were asked to evaluate the text on two fronts; Comprehensibility, which measures how easily the MT output is understood by the user, and Post-Editability, which indicates how much effort they *thought* would be necessary to refine the MT output to a publishable level. Based on a comparison of scores from human and automatic evaluations, they concluded that whereas automatic evaluation scores correlate with Comprehensibility scores, they do not correlate well with Post-Editability scores. However, this was a rather impressionistic evaluation and may well be worth examining by comparing with the actual amount of PE effort.

There have also been efforts to predict MT output quality automatically and indicate it by scores, which are collectively called a ‘confidence index’ or ‘confidence measure’. Some of these indices focus only on ST characteristics, such as sentence length, complexity, terminology coverage by the UD, and so on, while others take into account both ST and TT, comparing various characteristics of both, including sentence length, construction, and so forth (Gamon et al. 2005, Rojas & Aikawa 2006, Soricut & Echihabi 2010). Some of the studies have PE effort in mind (Specia et al. 2009a, Specia et al. 2009b, González-Rubio et al. 2010), but the usefulness of such scores has not been tested against actual PE effort as far as the author of the present study is aware.

2.4.2 ST characteristics and PE effort

Krings (2001) drew attention to the importance of examining the influence of the ST on the amount of PE effort. He investigated the effect of ST characteristics from a number of view points, including length, function, topic, degree of specialisation, structure, language pair, vocabulary, difficulty, and syntax, and found that some of the ST characteristics have influence not only on the ST-related process, such as reading and understanding, but also on TT production and evaluation. However, the unit of ST characteristics he employed was ‘text,’ and only the overall characteristics of texts were considered; for example, French imposes more effort intensiveness than English does, or longer text enables post-editors to familiarise themselves with the contents, thus speeding up the editing process. He did not examine ST characteristics at finer levels, such as sentences.

One way of controlling the influence of ST on MT quality is the use of controlled language (CL) rules. According to the definition of Nyberg et al. (2003: p245), CL is “an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style” whereby ambiguity and complexity of the texts are reduced. Bernth & Gdaniec (2001) coined the word ‘MTranslatability’ while suggesting rules for authoring texts that are more easily translated by MT systems than arbitrary writing. They suggested a number of writing rules in terms of grammar, ambiguity, style, punctuation, and spelling. To date, many international organisations have reported on the planning, building, and implementing of CL rules, and in most cases CL was said to be beneficial (Newton 1992, Mitamura & Nyberg 1995, Douglas & Hurst 1996, Nyberg & Mitamura 1996, Mitamura 1999, Rychtycky 2006). These studies have shown the effect of CL in terms of MT quality, but the question still remains as to the effectiveness of CL in an entire MT workflow, especially at the PE stage.

2.4.2.1 Effect of CL on PE effort

Allen (2003) noted that CL and PE approaches were often used in combination so that translation quality was improved, and consequently PE time became shorter. Perhaps CL could be regarded as, so to speak, a preventative effort that helps to reduce the need of a curative effort, PE. This makes more sense when translation is to be done from one language into many languages, since CL rules need to be applied only once to the

source language whereas PE needs to be done on all the target languages separately. But to what extent could PE effort be reduced by applying CL rules?

O'Brien (2007) focused on Negative Translatability Indicators (NTI), the characteristics of the ST that may impose difficulties in MT, and found those factors affect PE processes. She identified a number of NTIs, such as gerund, ungrammatical construct, and long noun phrase, and found that, although each of them affects PE speed at a different level, the presence of more NTIs generally slows down PE compared to the sentences that contain no or fewer NTIs. This indicates that controlling the input to MT may help to reduce temporal PE effort.

O'Brien & Roturier (2007) compared the impact of different CL rules on comprehensibility and post-editability of MT output. They have found that some CL rules, including avoidance of misspelling or sentences longer than 25 words, have a high impact on improving both comprehensibility and post-editability, and some CL rules, including the avoidance of parentheses, have a low impact on improving both. However, they also found that some CL rules, for example, restriction of the use of noun clusters and relative pronouns, have contrasting impacts on comprehensibility and post-editability of MT output. This may mean that it should not be assumed that the better quality MT output automatically results in lesser PE effort. This also supports the finding of Krings (2001) and Ramirez & Haller (2005) discussed in section 2.4.1.

There have been similar research efforts on the effect of CL on PE of SMT output. Aikawa et al. (2007) studied the impact of CL on MT output and PE effort. They showed that certain categories of CL rules, such as the use of formal style, elimination of ambiguity, and ensuring correct spelling have a significant impact on producing better MT output for all four target languages they investigated, namely, Arabic, Chinese, French, and Dutch. They also compared the textual differences of post-edited text against two MT outputs: one from the uncontrolled original ST and another from CL-applied ST. They claimed that the MT output from the CL rule-applied ST required less PE effort. However, they measured the difference of PE effort solely by examining the textual difference between MT output and the post-edited final product.

More recently, Temnikova & Orasan (2009) tested the effectiveness of CL rules defined for emergency-related text on human translation and post-editing of machine translation. They used seven target languages: Bulgarian, Slovenian, Russian, Spanish, Dutch, Maltese, and Greek, and the Google Translation Engine to machine translate the text. The effect was measured by time and the amount of editing. Their findings suggest that the machine translation of controlled text is generally faster to post-edit than that of uncontrolled text, but the effect of CL is much larger on human translation than machine translation.

2.4.3 PE typology

In order to understand the PE effort in detail, it is crucial to breakdown the overall PE effort into sub-processes and categorise them. There have been a number of PE typologies suggested (Laurian 1984, Allen 2003, Schäfer 2003, National Institute of Standards and Technology Information Access Division / Speech Group And Linguistic Data Consortium 2007, LISA), and most of them are linked to MT errors. For example, Allen (2003: p.307) introduces the Society for Automotive Engineering J2450 standard metric for translation quality, as a PE guideline employed at General Motors, which consists of:

- A. Wrong term
- B. Syntactic error
- C. Omission
- D. Word-structure or agreement error
- E. Misspelling
- F. Punctuation error
- G. Miscellaneous error

Schäfer (2003), in reporting ongoing PE guideline development in SAP, lists more finely defined MT errors that post-editors should address:

- 1. Lexical errors
 - 1.1 General vocabulary
 - 1.1.1 Function words
 - 1.1.2 Other categories

- 1.2 Terminology
- 1.3 Homographs / polysemic words
- 1.4 Idioms
- 2. Syntactic errors
 - 2.1 Sentence / clause analysis
 - 2.2 Syntagmatic structures
 - 2.3 Word order
- 3. Grammatical mistakes
 - 3.1 Tense
 - 3.2 Number
 - 3.3 Active / passive voice
- 4. Errors due to defective input text

A more widely employed error classification, especially in the IT industry, is LISA's QA Model:

- 1. Mistranslation
- 2. Accuracy
- 3. Terminology
- 4. Language
- 5. Country
- 6. Consistency

These categorisations can be helpful to create guidelines and evaluate PE results in commercial settings and also have been used for research purposes (Guerberof 2008, de Almeida & O'Brien 2010). However, in order to utilise these classification systems, one first must identify 'errors' and 'inappropriateness' in MT output that post-editors are supposed to correct, which adds an extra step that might introduce researchers' subjectivity. In order to understand what is happening in the actual PE effort in a non-preconceiving manner, we may need a method for harvesting the patterns by observing the editing process. Abekawa & Kageura (2008a), though they studied the revision process of human translation of English to Japanese as opposed to that of MT output, gleaned the actual ingredients of the modification process from the translated and edited text data. They looked at each modification and identified its 'primitive operation' (insertion, deletion, replacement, and transposition), then 'linguistic operation' (change of verb, particle, voice, etc.), 'reason' (awkward, mistranslation, etc.), and 'aim' (raise fluency, add information, etc.). This type of pattern identification approach may better help us to understand more about what is actually going on in human editing processes in an open-ended manner, than classifying each sub-process into pre-defined categories.

Temnikova (2010) proposed another category set of MT errors that require PE, ranked by the presumed cognitive PE effort intensiveness. Based on the cognitive model of reading, working memory theory, and error detection studies, she suggested ten categories, in the order from least to most effort intensive tasks:

1. Correct word, incorrect form
2. Incorrect style synonym
3. Incorrect word
4. Extra word
5. Missing word
6. Idiomatic expression
7. Wrong Punctuation
8. Missing Punctuation
9. Word Order at Word level
10. Word Order at Phrase level

She used these categories in order to estimate the amount of cognitive PE effort based on the number of corrections made during PE, though the appropriateness of the ranking is not tested against any empirical data.

2.4.4 Post-editor variance

As with any human activities, there must be individual variations in handling PE tasks. O'Brien (2006b), Guerberof (2008), and Plitt & Masselot (2010) have all reported the speed difference between post-editors in their studies; the speed of the fastest post-editors were from 190% to over 230% of that of the slowest post-editors. Krings pointed out such variations can be caused by their competence as a translator, the personality, familiarity with MT programs, computers in general, word processing programs, and so on (Krings 2001). These factors, along with other possible factors, such as experience in PE and expertise in the subject domain, may need to be taken into account in analysing and drawing inference from the data.

2.4.5 Research settings

Obtaining scientific evidence as to real-life PE effort is a challenging task. Hiring professionals for long hours may be costly. Also, experimental environments need to be controlled to some extent but they also have to be close to the normal work environment.

Krings (2001) recruited in his study 16 ‘semi-professional’ subjects, who are professionally trained but not experienced technical translators, and the number of words in the ST in three languages he tested, namely, English, French, and German, were 1,352, 762, and 325 respectively. He needed to limit the size of the text partly because of the method he employed, TAP, which requires a considerable amount of time and effort to analyse. The TAP also might have compromised the ability to reproduce the normal work environment as the subjects had to verbalise all thoughts they had during PE, which, for most people, is an unusual way of working.

O’Brien (2006b) emphasises the importance of using professionals as subjects in order to obtain authentic data. She recruited twelve professional translators in her study, and had them either translate or post-edit the text translated from 1,777 words of English text.

Guerberof (2008) conducted a pilot study with nine professional translators, and the length of the text she used was 791 words. She compared the productivity between translation from scratch, editing of TM fuzzy match segments, and post-editing of MT output, and the total number of words for each test set was 265, 264, and 262, respectively. The participant post-editors were required to use text editing software devised for the study, which might have compromised the reproduction of a normal work environment.

The research conducted by Plitt & Masselot (2010) probably is the first large-scale study of professional post-editors. They employed a total of twelve professional post-editors of FIGS, and the cross-product of the text either translated or post-edited by 12 participants for all four languages amounted to 144,648 source words in total. As for the working environment, however, they developed a special text editing software in order

to precisely record the editing time. This, similar to Guerberof's study mentioned above, might have had an influence on the post-editors' work.

2.4.6 Analysis method

As mentioned above, there have been a number of studies that have strived to analyse PE effort empirically. However, quantitative analysis methods are generally limited to either descriptive statistics, or inferential statistics based on a single explanatory variable, which has a limited power of controlling conditions. As with any human activity, PE is a complex activity, which is considered to be affected by a number of conditions at the same time. More sophisticated statistical techniques may help to take into account multiple explanatory variables and provide deeper insight into the PE effort.

2.5 MT and TM

A TM is a tool that facilitates human translation, and it generally includes a database that stores the ST and the translation of the text, and some interfaces for referencing past translation and terminology, and for editing the text. The database can be built either retrospectively, from a parallel corpus of the already translated ST and TT, or interactively, on a sentence-by-sentence basis as a translator translates the ST into TT. The database stores the text in 'translation units', aligned pairs of ST and TT segments, which is often a sentence or any 'segment' delimited by defined symbols (period, hard return, colon, etc.). When a translator translates a segment, the TM software searches the database for the past translation of either exactly the same or partly the same ST. The former is called an 'exact match' and the latter is called a 'fuzzy match.' TM has become a popular tool for speeding up commercial translation since the late 1990s (Bowker 2002). Technical documents and instruction manuals, which tend to have many repetitions of the same or similar sentences, and are subject to relatively frequent revision, can especially benefit from TM.

2.5.1 Combining MT and TM

Bowker (ibid) reported in 2002, that some developers started to combine TM and MT, but such applications have relatively recently become popular; SDL Trados 2007 is equipped with the function to allow the users to access their MT system, Multicorpora

offers a Systran plug-in, and Wordfast and Swordfish provide access to web-based MT engines including Google Translate (Garcia 2007). Apart from such integration, some users of TM and MT combine the two technologies together on their own initiatives; they pre-translate the text with TM exact and fuzzy matches, and machine translate only the sentences that do not have corresponding translation in the TM database (O'Brien 2002, Dillinger & Lommel 2004). TAUS reports that post-editing is now mostly done on such "hybrid TM + MT documents" (TAUS 2010: p.9).

This has raised the issue of productivity and payment for post-editors: 'Which is more productive, post-editing MT output or editing TM fuzzy matches?', which in turn, can be a decisive factor in setting the price for PE of MT and editing of TM. Guerberof's study focuses on this matter, and the results suggested that, in the case of English to Spanish translation, PE of MT output is faster than editing of TM 80-90% fuzzy matches (Guerberof 2008). However, she also found that the post-editor variance was high.

2.6 MT and PE research in Japan

There has been a wealth of MT-related research in Japan, since "The Japanese are amongst the world's leaders in developing machine translation" as Kondo & Wakabayashi state (2009: p.476). The first Japanese MT system, Yamato, was developed in 1959 (Takahashi et al. 2003), which was an MT-devoted machine that translated from English to Japanese. In the 1970s, the Japanese MT authority Makoto Nagao started developing an MT system that translated the titles of scientific and engineering papers, focusing on their relatively limited structural patterns (Nagao et al. 1982). He reported that this system produced correct translation about 80% of the time. In 1981, he then proposed the world's first Example-Based MT system (Nagao 1984). The first commercial Japanese MT system was introduced in the mid 1980s by Fujitsu.⁸ In 1989, Japan Broadcasting Corporation (NHK: Nippon Hoso Kyokai) started a test run for broadcasting machine-translated news from abroad (Aizawa et al. 1990). Today, there are numerous commercial MT systems that handle Japanese as a source and/or

⁸ <http://software.fujitsu.com/jp/atlas/> [Last accessed: 19/10/2010]

target language, most of which are for general use, but some for specific domains, such as medical text.

The area of focus for Japanese MT research and development spans across a broad range. There have been efforts for developing community based online MT systems, such as Yakushite-net,⁹ where people from different domains form communities, and maintain domain specific dictionaries by collaborative effort. Human and automatic evaluation of MT system and MT output quality also has been researched widely. The Asia-Pacific Association for Machine Translation (AAMT) has a working group that is specialised in constructing and publicising evaluation test sets both for MT vendors and users, and establishing evaluation methods and indices. The Japan Electronic Industry Development Association (JEIDA) proposed three novel human evaluation methods in 1991: one for potential MT users to evaluate MT systems in terms of economical effectiveness, another also for users to compare the technical competence between different MT systems, and a third for MT researchers and developers to evaluate the technical level of a developed MT system. The major advantages of these methods include: a) the evaluation can be done following the provided check list which addresses each issue to be checked in an isolated and simple manner, and are thus easily and objectively measured, b) the result of the measurement is represented as scores, thus can be numerically rated, and c) the scores can be converted into a radar chart, which makes the overall judgement easy and user friendly (Nomura & Isahara 1992). JEIDA proposed a revised version in 1993, which was enhanced by inclusion of translation quality evaluation (Nomura 1993, Isahara 1995). Other evaluation-related research projects include a proposal of MT evaluation from three aspects, namely, the quality of the MT system, the dictionary, and the ST (Kiuchi & Kaihara 1991), a proposal for a new AEM (Ehara 2007a), and a meta-evaluation study of AEMs based on patent documents (Echizen-ya et al. 2009). The MT-related research has been especially accelerated in the area of patent translation (Oshio 1980, Neumann 2005, Fujii et al. 2009), where Japanese to English is the main language pair, and the difficulty lies in parsing long and complicated Japanese sentences that are specific to patent texts (Yokoyama & Kennendai 2007). A large amount of MT-related research and development in Japan has been conducted in the Japanese-English language pair, into

⁹ <http://www.yakushite.net> [Last accessed: 19/10/2010]

both directions, but there have been initiatives that involve other languages, especially Asian languages, such as the MMT (Multi-lingual Machine Translation) project in five Asian languages, namely, Japanese, Chinese, Indonesian, Malaysian, and Thai (Funaki 1993), and other research projects in Japanese-Korean (Makita et al. 2003, Song & Bond 2009) and Japanese-Chinese (Sumita et al. 2007, Isahara et al. 2007).

Japanese CL has also been researched, and CL research has become especially active since 2007 in the area of patent documents, and recently an organisation called Technical Japanese Association¹⁰ was founded (Watanabe 2010). Its aim is to develop a new Japanese language framework in order to facilitate technical information sharing both in human-to-human communication and computerised/automated text processing, such as authoring, search, and translation. The application of technical Japanese is expected to improve the MT output quality of technical text (Kumano 2008), and automatic paraphrasing from non-controlled Japanese to controlled Japanese is also being studied (Kumano et al. 2009). There have also been studies on automatic pre-editing (Yoshimi et al. 2000, Yoshimi 2001) and post-editing (Yamamoto 1999, Ehara 2007b). However, although the necessity of human PE has been recognised for a long time in Japan (Itagaki et al. 1999), it has largely been neglected as a field of MT research, except for some development effort for human PE support environment (Usui et al. 1986). Also, MT research in the context of the IT industry has been sparse to this day, even though computer manuals had been considered suitable content for translation by MT from an early stage of MT development in Japan (Nagao et al. 1980).

Only very recently, human PE is attracting attention. Yamada (2010) conducted an experiment to measure the effect of CL rules on the PE workload. He recruited eight Masters level translation and interpretation students as subjects, and had them post-edit the MT output of technical manuals translated from English to Japanese by Google Translate (an SMT system) both from CL-applied and non-CL-applied ST. The results showed that, on average, the amount of necessary textual change was reduced by one third when using CL-applied ST compared to non-CL-applied ST. He also examined participants' impressions about the workload compared to translating from scratch. Their answers suggest that, on average, the participants felt that the workload of post-editing the MT output translated from non-CL-applied ST was about 87% of that of

¹⁰ <http://www.tech-jpn.jp> [Last accessed: 20/10/2010]

translating from scratch, while the workload of post-editing the MT output translated from CL-applied ST was about 73% of that of translating from scratch. These results show the potential in MT + PE workflow in the context of technical translation, though there still is much room for research and improvement.

2.7 Formulation of Research Questions and Research Plans

As stated in the Introduction, the fundamental research question of this study is:

What determines the amount of post-editing effort?

which is too broad to be answered directly, and thus needs to be divided into more specific and measurable research questions. The findings from the literature review give us important insights to help achieve this.

Firstly, research into the relationship between the amount of effort in PE process and the amount of necessary edits (which correlates with MT output quality) indicates that they do not always have a linear relationship. However, the level of correlation is not quantified in detail, and it has not been made clear what causes the variance between them. Secondly, characteristics of ST have proved to have some impact on PE effort, but the prior research has been done mostly from the viewpoint of controlled language, and there still may be various elements that affect the amount of PE effort. Thirdly, PE typologies have been attempted mostly from the viewpoint of quality assurance and evaluation, and there has not been an extensive effort to identify PE patterns inductively by observing actual PE activities. Also, as integration of MT and TM becomes more and more prevalent, there is a need for understanding the relationship between post-editing of MT and editing of TM in more detail. The research context is also important. In order to understand the PE effort in an authentic situation, a study needs to be conducted in a condition that reproduces professionals' normal work environment as much as possible. In addition, some sophisticated statistical methods may need to be employed so that the findings can be quantified taking into consideration a number of conditions involving the PE environment.

Consideration of these implications have led to the formulation of four core research questions:

Core Research Questions (Quantitative Questions):

RQ1: How does the amount of editing correlate with the amount of effort in post-editing of English to Japanese MT output?

RQ2: What characteristics of the English source text have significant influence on the amount of PE effort irrespective of individual traits of post-editors?

RQ3: What types of PE operation have significant influence on the amount of PE effort irrespective of individual traits of post-editors?

RQ4: What are the differences between editing of TM match segments and post-editing of MT output?

While these questions need to be answered by quantifying the results, understanding human activity inevitably involves ‘how’ questions. In order to take into account the case-by-case difference of the situation and individual differences of the post-editors, we add a further question:

Further Inquiry (Qualitative Question):

RQ5: How do different attributes, techniques, and/or behaviour of post-editors affect PE practice?

In addressing each of these questions, it is hoped that this study will provide us with knowledge that helps to more efficiently integrate and streamline the PE process in the entire translation production workflow that employs MT. By identifying the most problematic problems in post-editing MT output, we may be able to gain understanding that helps to improve the MT engine itself or develop automated PE technology, both of which will help unload some of the PE tasks from post-editors. On the other hand, by identifying the issues in performing PE, we may be able to suggest ways to improve PE training, PE guidelines, and the MT workflow, all of which may help to make PE easier.

2.8 Concluding Remarks

In this chapter, we first defined ‘post-editing’ by reviewing some of the prior definitions of the same. After that we reviewed relevant research effort to date, focusing first on 1) PE strategies and methodologies, and then 2) measuring methods of the amount of PE effort and 3) possible factors that affect the amount of PE effort. We then, in introducing some of the main findings, pointed out research gaps. Some are related to research findings, and others are related to research methodologies. From these implications, we formulated the research questions of the present study.

PART II

CONDUCTING THE STUDY

Chapter 3 RESEARCH METHODOLOGY

The purpose of this chapter is to explain the overall theoretical methodological approach taken to answer the research questions of this study. Section 3.1 discusses a suitable research framework. Sections 3.2 and 3.3 explain the research design in detail, following which the conceptual definitions used in the research questions will be operationalised in section 3.4. Section 3.5 discusses the statistical methods employed in the present study, and section 3.6 discusses the quality assessment issue.

After this chapter, Chapter 4 reports on the preliminary study, following which Chapter 5 explains practical methods for data collection and analysis, that have been refined based on the lessons learned from the preliminary study.

3.1 Research Framework

Hughes & Hayhoe (2007) suggest that the goals of research guide the researchers to a suitable research methodology. Among the list of research goals they explain, namely, Theoretical, Empirical, Interpretivist, Postmodern, Developmental, and Evaluative (ibid: p.7), *Empirical* and *Interpretivist* research goals are relevant to the present study. Empirical research secures objectivity by showing the results in numbers, while interpretive research takes an open-ended approach, for instance, instead of asking ‘*Is A better than B?*’, an interpretive research asks ‘*What makes A better?*’. In the present study, RQ1, RQ2, RQ3, and RQ4 (see pp. 31-31) are empirical, with RQ1 investigating a cause and effect relationship, RQ2 and RQ3 examining the effect of each variable, while RQ4 compares the results from two groups. However, RQ3 also requires first an interpretive approach in order to describe or explore what affects the amount of PE effort in an open-ended manner, before quantifying the magnitude of each issue. RQ5 will need to be addressed by interpretive approaches in order to answer *how* questions.

As Hughes & Hayhoe (ibid) note, research goals and research methods are strongly associated. Empirical research is associated with quantitative methods, while interpretive research with qualitative methods. However, having qualitative methods and quantitative methods independent of each other may not be sufficient in answering

the research questions in the present study, in which qualitatively interpretable data, such as post-edited text, and purely quantitative data, such as quantified PE effort, need to be analysed in combination. For this reason, the author of the present study chose to employ a *mixed methods research* design.

3.2 Mixed Methods Research Design

According to Creswell & Plano Clark (2007), mixed methods research design, while having the advantage of combining quantitative and qualitative research and facilitating better understanding of research problems, is not merely a combination of qualitative and quantitative research, but provides more comprehensive evidence and gives an opportunity to look at the problems from both inductive and deductive points of view. Creswell & Plano Clark suggest four major mixed methods designs, namely, *Triangulation Design*, which uses complementary data about the same phenomena to confirm the findings, *Embedded Design*, in which the secondary analysis is embedded in the primary analysis and provides a supportive role, *Explanatory Design*, in which quantitative results are further explained by qualitative investigation, and *Exploratory Design*, in which the findings from the first, qualitative phase inform the methods of the second, quantitative phase (ibid: p.62-79). Creswell & Plano Clark encourage the researchers of mixed methods research to choose a single design and follow the framework and logic of it for making the research manageable. However, in order to best address the research questions in the present study, it is desirable to have both an exploratory and an explanatory phase, and thus we chose to combine Exploratory and Explanatory designs in a cascaded manner. Figure 3.1 depicts the overall design adopted in this study.

The National Research Council of the USA suggests three types of research inquiries (National Research Council 2002: p.97-126): ‘*What is happening?*’ (description), ‘*Is there a systematic effect?*’ (cause), and ‘*How is it happening?*’ (process or mechanism). The research questions in the present study include all these three types of inquiries; firstly, we explore and find out what is going on during PE in [2], secondly, we quantify the magnitude of each finding in [4], and finally, we try to explain how and why they happen in [6].

3.2.1 ‘What is happening?’

Our central interests (RQ2 and RQ3) require first clarification of the ingredients of ‘ST characteristics’ and ‘PE operation types’. The ST characteristics that have impact on PE have been suggested by a number of researchers. In the present study, possible candidates of ST characteristics have been identified through the review of such related works and findings from the preliminary study, which will be discussed in Chapter 4. On the other hand, identifying PE operation types will be done by first qualitatively comparing MT output and PE output. This procedure fits the ‘Taxonomy Development’ model of Exploratory design Creswell & Plano Clark propose, in which “the initial qualitative phase is conducted to identify important variables, develop a taxonomy or classification system, or develop an emergent theory, and the secondary, quantitative phase tests or studies these results in more detail” (Creswell & Plano Clark 2007: p.77). Then, identified ST characteristics and PE operation types will be quantified to check their significance. The process of PE operation type identification will be discussed in detail in section 3.4.4.

3.2.2 ‘Is there a systematic effect?’

The main goal of the present study is to quantify the effect of different variables on PE effort. This will be built on the theories from related works and descriptions from the data collected for the present study, as mentioned in section 3.2.1. This approach meets the requirement of a causal work as described by the National Research Council of the USA (2002: p.108): “ideally, a strong theoretical base as well as extensive descriptive information are in place to provide the intellectual foundation for understanding causal relationships”. While qualitative findings are of importance in their own right,

quantifying such findings will help in prioritising the issues for improving MT output, streamlining MT workflow, and developing PE guidelines and training materials.

3.2.3 ‘How is it happening?’

Some of the quantitative findings may be self-explanatory, but other quantitative findings will be further explained by qualitatively examining their meaning, especially for somewhat unexpected findings. Creswell & Plano Clark define the ‘Follow-up Explanatory Model’ as a two-phase design in which the researcher identifies specific quantitative results that need further qualitative interpretation as to what they mean (Creswell & Plano Clark 2007). In the present study, case by case investigation will be conducted on some of the unexpected or specifically significant findings, which may help in clarifying the mechanism of such phenomena.

3.3 Observational Experiment (Data Collection Phase)

In addressing these questions, we put a strong emphasis on collecting the data from a real-life, authentic setting, which includes a) observing professionals’ work, b) in their normal work environment, c) using real data (text), and d) in a manner that is as non-invasive as possible.

In order to fulfil these prerequisites, we need to:

- a) Hire professionals as participants, and have them post-edit the text as they normally do
- b) Conduct the study in an environment that is familiar to participants
- c) Use data from the industry that employs a state-of-the-art MT workflow
- d) Use tools and software that are familiar to professionals and that do not interrupt normal PE process

These conditions make it difficult, if not impossible, to conduct an experiment in a highly controlled environment, such as in a laboratory, with all conditions and variables manipulated by the researcher. However, as Frey points out, “It is not always possible or even desirable, to manipulate an independent variable” (Frey et al. 1991: p.158), in which case less control is exercised, and naturally occurring conditions can be

employed as variables, which are called “observed variables” (ibid: p.158). The present study, therefore, will largely be based on observed variables, in an *observational experimental design*, in which data will be gathered in a real-life situation, but in a semi-controlled manner (that is, have post-editors work on certain documents under specified guidelines, rather than observing each post-editor’s own daily work).

3.4 Operationalisation

In order to quantitatively examine the phenomena, conceptual definitions introduced in the research questions need to be converted into measurable variables, that is, operationalisation. Operationalisation is the process in which *observable characteristics* of the concepts to be measured are determined (Frey et al. 1991). Operationalisation is also necessary when even though there is a widely agreed upon theoretical construct, how to measure it is debatable (Boslaugh & Watters 2008). The most important thing to keep in mind in the process of operationalisation is to ensure the consistency of the meaning between conceptual definitions and operational definitions, though it is impossible to capture all the conceptual definitions (Frey et al. 1991). By making sure that the operationalised variable captures what is intended to be measured in the study, the researcher can manage ‘internal validity’ (Hughes & Hayhoe 2007), which enables us to draw valid conclusions (Rasinger 2008).

The following subsections explain how each of the conceptual definitions in research questions have been operationalised keeping in mind internal validity.

3.4.1 Amount of PE effort (dependent variable)

Throughout RQ1 to RQ4, the amount of PE effort will be used as the dependent variable. As discussed in Chapter 2, Krings (2001) suggested the need to distinguish three kinds of PE effort: cognitive, temporal, and technical. Among them, we decided to employ temporal effort as a primary measurement of the amount of PE effort.

There are a number of benefits to taking this approach especially in the commercial context. First of all, temporal effort is the most externally visible aspect of PE effort, and can be measured easily and objectively (Krings ibid). Secondly, time recording is

not a very complex function thus can be relatively easily embedded in post-editors' standard work environments, which makes it a non-invasive technique and enables the capture of real-life data rather than conducting the experiment in a lab. Thirdly, time is a simple numerical measure, and a large amount of data can be processed and analysed with relative ease. More importantly, time is an objective measure that is easily quantified. Such amenability to quantification is important to the industry as a means of performance measurement. Finally, time affects the production schedule and cost, and is therefore relevant to the industry.

Krings outlined the relationship between the three effort types as: 1) the determining variable of the temporal effort is the cognitive effort, that is, the more demanding the cognitive task is, the longer the task takes to process, and although 2) the technical effort is guided by the cognitive effort, that is, the decisions for an action are made in the cognitive process, 3) some technical efforts involve more complicated actions, such as a number of word reordering, which may require more temporal effort. Although temporal PE effort alone does not perfectly represent the overall PE effort, these remarks of Krings altogether may indicate that the difficulty of the cognitive process combined with the difficulty of the technical process determines the amount of temporal effort. As cognitive and technical efforts often occur simultaneously (start typing while still thinking, or rethink while typing, etc), the temporal effort ultimately can be thought to represent the sum of the cognitive and technical effort but with some overlaps. Figure 3.2 is a simplified diagram of the concept of PE effort employed in this study.

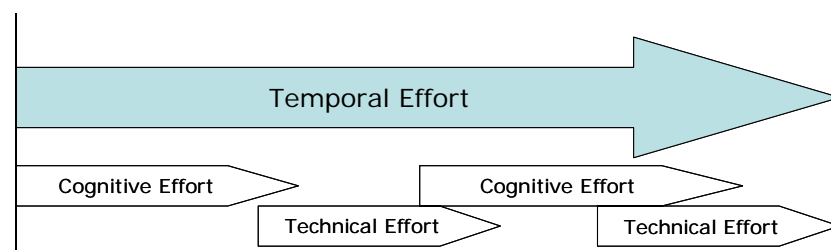


Figure 3.2 Concept of PE effort

However, it may not be appropriate to directly compare the PE time of a sentence to that of another sentence with a different length, as the amount of text should have some effect on the processing time. Therefore, it is appropriate to normalise the time by the

length of the sentence, producing the PE *speed*. Krings also employed speed and stated that the “speed provides significantly more information than the absolute processing time” (ibid: p.277). Although normalisation can be done by using either the length of the ST or TT, we chose to use the ST length, as it makes it possible for us to directly compare the post-edited output between different post-editors, who may have produced TT of different length from the same ST. In addition, among a number of possible units of the speed, such as words per minute, words per second, minutes per word, and seconds per word, we decided to use words per minute, which enables us to keep in line with prior research in the same field (Krings 2001, O’Brien 2006b, Guerberof 2008). The actual technique for measuring the time and the word count to obtain the speed will be discussed in detail in section 5.1.1.

3.4.2 Amount of editing (independent variable)

Quantifying the amount of editing may not be a straightforward task. For example, compare the following two sentences.

Then you can specify the target name in your search criteria.

You cannot add the target name in your search criteria box.

The word number has been increased in the second one (11 for the first, and 12 for the second) but the character number has been decreased (51 and 49). A word has been deleted in the second sentence (‘Then’), two words have been added (‘not’ and ‘box’), a word has been replaced by another (‘specify’ to ‘add’). If we are to take into account the impact of changes, we need to consider the fact that while the change from ‘specify’ to ‘add’ does not alter the fundamental meaning of the sentence, addition of ‘not’ does. There can be numerous ways of defining and measuring the ‘amount of editing’.

In order to avoid such subjectivity, we employ Krings’s idea of measuring the surface similarity between the two texts, disregarding the semantic aspect (2001). He devised a method of measuring textual similarity, that is, counting the number of words common to MT output and PE output disregarding the word order, and dividing it by the number of ST words. If the same word was used with different inflection, it was counted with a factor of 0.5. He notes that textual similarity represents the extent of post-editors’

intervention on the text, which is an “obvious method of determining PE effort” (ibid: p.531). The advantages of this method, he argues, are a) it provides a satisfactory level of validity, in that it gives the scores that “agree with our intuitive language sense” (ibid: p.296), b) it is highly operational, in that it is free from subjective interpretation of similarity, and c) it gives the possibility of future automation, although at the time of his publication no automation effort was being made.

Fortunately, after nearly a decade, we now have numerous computer programs that do similar jobs as his method. They are usually collectively called ‘automatic evaluation metrics’ (AEM), as they have originally been developed to automatically evaluate the quality of MT output by comparing it with one or more ‘correct translation(s)’ (also known as ‘reference translations’). In the present study, however, we employ these metrics solely for the purpose of comparing MT output and PE results in order to quantify the amount of editing performed during PE, and *not* for evaluating the MT output quality.

Different AEMs have been designed based on different viewpoints about the ‘closeness’ between two or more texts, and use different algorithms to compute the scores. Among a number of available AEMs, we chose to test three widely used AEMs, namely, BLEU (Papineni et al. 2002), GTM (Melamed et al. 2003, Turian et al. 2003), and TER (Snover et al. 2006). The following are brief descriptions of the AEMs employed.

BLEU (BiLingual Evaluation Understudy) (Papineni et al. 2002) counts the number of words found in common in two texts, and uses precision to compute the level of similarity, rewarding correct word order by double-counting the 2, 3, and 4-word exact matches. The BLEU scores range from 0 to 1; if two sentences are identical, the score is 1, if two have no word in common, 0. The version used in this study was v11b.¹¹

GTM (General Text Matcher), (Melamed et al. 2003, Turian et al. 2003) also counts the number of overlapping words between two texts, but uses not only precision but also recall to compute the similarity level. The importance of the word order can be adjusted by specifying options. The GTM scores range from 0 to 1; As with BLEU, if two

¹¹ <http://www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html> [Last accessed: 22/2/2010]

sentences are identical, the score is 1, if two have no word in common, 0. The version used in this study was 1.3.¹²

TER (Translation Edit Rate) (Snover et al. 2006), on the other hand, counts the minimum necessary to transform the first text to the second, rather than counting the number of *matches* as BLEU and GTM do. It counts the number of insertions, deletions, and substitutions of words, and shifts of sequences of words, and divides it by the number of words in the second text. This gives the perfect match a score of 0 (0 edits needed), as opposed to BLEU and GTM. Moreover, since the lengths of the two sentences can be infinitely different, there is no limit for the upper score. The version used in this study was 0.7.25.¹³

The reasons for choosing these metrics were: 1) they are applicable for both a wide variety of European languages and Japanese, and 2) they are commonly used in relevant research and literature. Though BLEU is not designed for sentence level evaluation, it was also employed in this study since BLEU is the de facto standard evaluation metric, it is tested also for sentence-level evaluation in literature (Callison-Burch et al. 2008), and comparison with PE speed is a rather new approach and thus it was deemed to be worth testing.

While these metrics measure the similarity or difference on a word-by-word basis, the Japanese writing system does not insert spaces to mark the boundary of words. Therefore, the Japanese text was divided or ‘tokenised’ by means of a Japanese tokeniser and morphoanalyser, MeCab, into words or ‘morphs’¹⁴ to be made available for processing by AEMs.

3.4.3 Characteristics of source text (independent variable)

As mentioned in section 2.4.2, Krings (2001) emphasised that the primary observation as to PE process in quantitative analysis is the influence of the ST on PE effort. He

¹² <http://nlp.cs.nyu.edu/GTM/> [Last accessed: 22/2/2010]

¹³ <http://www.cs.umd.edu/~snover/tercom/> [Last accessed: 22/2/2010]

¹⁴ In the present study, the word ‘morph’ is used to refer to Japanese equivalent of ‘word’. Detailed discussions on morph and tokenisers can be found in 5.1.2.1.

considered several characteristics, including language pairs, topic types, comprehensibility, function of the text, and length of the text, and suggested that the ST has a strong impact on PE processes. O'Brien (2005, 2006b, 2007) focused on Negative Translatability Indicators (NTI), the characteristics of the ST that may impose difficulties in translation, and controlled language rules, the rules according to which the ST is written in order to make the text easier for MT systems to translate, and found such factors affect PE processes. There can also be a number of other different aspects that can describe the 'characteristics of ST'. In this study, some of the known explanatory variables from the past research and some other variables that have been found significant from the preliminary study will be employed. Identifying ST characteristics that influence the amount of PE effort is particularly beneficial since such factors can improve the accuracy in predicting the amount of PE effort based on the ST.

3.4.4 Types of PE operation (independent variable)

As discussed in section 2.4.3, there already are a number of PE typologies suggested and used by researchers and organisations. However, it was decided that we would extract the PE operation patterns by comparing the actual MT output and PE output obtained in the present study for the following reasons. As mentioned in section 2.4.3, classification according to existing typologies, since they are linked to MT errors or recommended PE exercises, involves assessing the accuracy or appropriateness of MT output and checking the PE operation against it in order to determine whether or not a given PE operation is appropriate, which adds an extra step in classification. More importantly, it does not suit the purpose of the present study; our focus is on observing what is going on during PE, and not assessing the post-editors' work.

Analysing qualitative data needs to be done "using increasing levels of abstraction" (Creswell & Plano Clark 2007: p.30), in other words, the data need to be coded. The first step in coding in the present study was to automatically classify the editing operation on a word by word basis, and identify the part of speech of each word, using a somewhat similar approach to Groves and Schmidtke's PE pattern identification, in which all PE operation patterns were extracted by tracing the minimum path of each edit made on the MT output to produce PE output (Groves & Schmidtke 2009, cf. section

2.3.3). This made it possible for us to start with unequivocal classification and avoid researcher's subjectivity. This will be discussed in more detail along with the coding results in section 6.2.1.

3.5 Statistical Data Analysis

Since the main focus of the present study is the quantitative analysis of the amount of PE effort, we need to employ suitable statistical methods. As a starting point for quantitative analyses, we use *descriptive statistics* using tables and graphs to summarise the relevant information about the data set. Then we move onto *inferential statistics* to understand the relationships and draw conclusions. Our main statistical techniques are Pearson correlation coefficients, for answering RQ1, and multiple linear regression (MLR), for answering RQ2, RQ3, and part of RQ4.

The correlation coefficient is a measure that examines the tendency of two variables to change their values together, that is, when one increases, the other increases or decreases. The value of a correlation coefficient ranges from -1 to 1, with zero meaning no relationship, and higher values meaning stronger positive (towards 1) or negative (towards -1) relationships. Although it shows the level of a correlation, that does not imply the causation; we cannot automatically assume that the change of one variable has caused that of another. In the case of the present study, however, we are comparing the textual changes and the resulting amount of PE effort (PE speed), thus it is reasonable to think that the amount of editing done in a sentence has determined the PE speed of that sentence, and not the other way around.

MLR is a useful technique to quantify the relationship between multiple independent variables and a dependent variable. This method makes it possible to observe the effect of more than one independent variable and improve the estimation of the values of dependent variables (Woods et al. 1986), by allowing us to “explicitly control for many other factors that simultaneously affect the dependent variable”, which enables us to examine the effect of each variable in “*ceteris paribus*”¹⁵ (Wooldridge 2006: p.73). We

¹⁵ [adverb] with other conditions remaining the same; other things being equal (Oxford Dictionary of English)

believe this fits the purpose of our study, where supposedly a number of conditions including various ST characteristics and PE operations may affect the amount of PE effort at the same time. The benefits of this method include that it can accommodate different types of data, such as numerical data and binary data, and it does not assume normal distributions of independent variables, which gives us flexibility in choosing the independent variables. The validity of the MLR estimation can be tested by means of a number of post-estimation tests, which will be presented in Appendix C.

3.6 MT/PE Quality Assessment

In the present study, the quality of MT output is not evaluated since we are interested in the relationship between certain ST characteristics or PE operation types and the amount of PE effort, and the relationship between MT output quality and the amount of PE effort is beyond the scope of our study. However, assuming that the participant post-editors understand the quality goal of PE in this study and perform PE accordingly, the amount of editing could be considered to have some indication of the quality of MT output. In addition, as discussed in 2.4.1, MT quality and the amount of editing have been found to have a linear relationship. Nevertheless, as MT quality cannot be equated solely with the amount of editing, we will not discuss MT quality issue in this study.

Similarly, the quality of post-edited text is not assessed either as we are focusing on understanding the PE activities performed under certain guidelines. PE quality assessment also lies outside the scope of our research questions, and could not be adequately addressed in the time available. However, it may present an opportunity for future research, such as, the relationship between different quality measurements and the amount of PE effort.

3.7 Concluding Remarks

This chapter focused on the philosophical framework and overall research design of the study. The chapter first considered which research framework was suitable, and presented a detailed design, in which a cascaded mixed methods design was introduced. It then explained the concept of ‘observational experiment’ which best addresses the

need for answering the research questions in the present study. Then it addressed operationalisation by transferring some of the conceptual definitions into measurable variables and also explaining the method to be taken in operationalising other variables at a later stage. It also explained the statistical methods employed in the present study. Finally, it discussed the quality assessment of MT output and PE result.

Chapter 4 PRELIMINARY STUDY

Having formulated the research questions and suitable research methodology, we conducted a preliminary study in order to finalise the methods. In this chapter, the objectives (section 4.1), methods (section 4.2), and findings (sections 4.3 and 4.4) of the preliminary study will be reported. Based on the lessons learned (section 4.5), methods for the main study have been refined, which will be explained in detail in Chapter 5.

4.1 Objectives

There were two main objectives for conducting the preliminary study. One was to obtain tentative answers to the core research questions to supplement the operationalisation process. The operationalisation of the concepts used for this study had partly been determined through the literature and completed by learning from the data, as will be discussed in 4.4.1, 4.4.2, and 4.4.3. Another objective is to be informed about methodological strategies. The proposed methods and devised tools were tested in order to ensure that they had no technical flaws that might impede the collection of data, which facilitated the finalisation of the methods for the main study project.

4.2 Methods

4.2.1 Preparation of test corpus

The experiment was conducted on a test corpus that contained 4,784 English words in 475 sentences in 33 files. The files were randomly selected from a Symantec corporation software manual, which was written to conform to Symantec's controlled language rules. The text was machine translated into Japanese using Systran version 5, using customised Symantec user dictionaries. It was also pre- and post-processed by Symantec's automatic processing scripts, which performed global search and replace to make the ST more amenable for MT, such as protection of XML tags, and to make as many repetitive corrections as possible on the target text before the human PE process, including the deletion of unnecessary spaces and replacement of some lexical items.

4.2.2 Post-editing session

The machine-translated Japanese text was post-edited by three Japanese native speaker translators, using SDL Trados Translator's Workbench (version 7), one of the most popular industry standard TM tools (Lagoudaki 2006), and TagEditor (version 7), a widely used translation/PE tool (TAUS 2010). Figure 4.1 shows an example of a typical use of these tools.

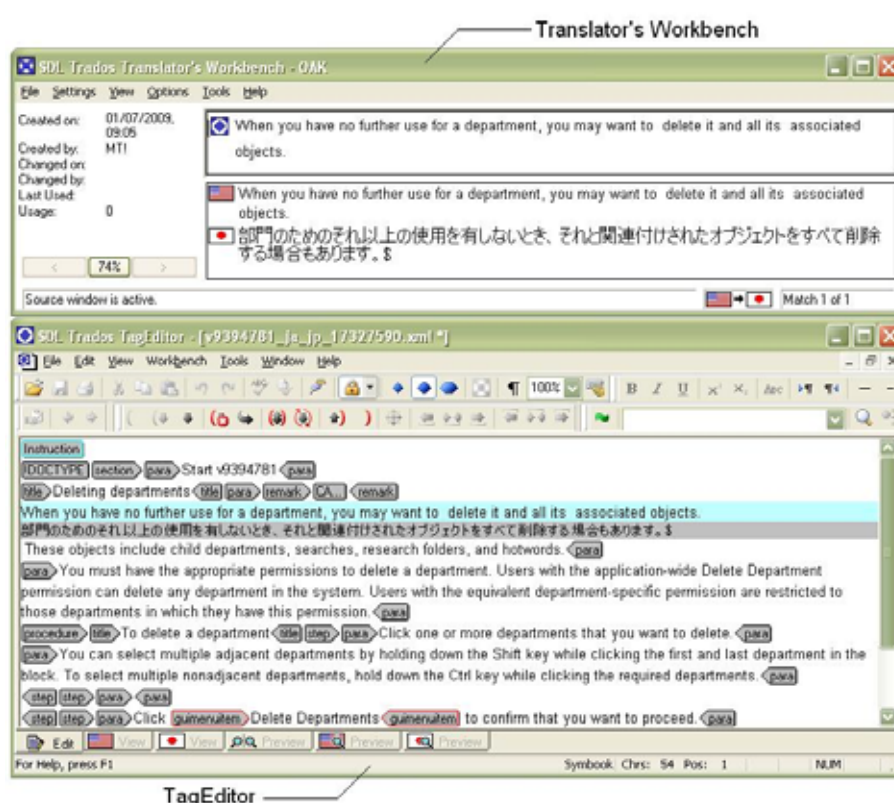


Figure 4.1 Trados tools

Translator's Workbench consists mainly of three parts: ST segment (upper right), which shows the source text currently being translated or edited, TM match segment (lower right), which shows the ST-TT pair found in the translation memory, and information fields (left), which shows various information about the TM match segment, including the dates of creation, change, and use, the author or editor of the TM match segment, and the fuzzy match ratio in per cent. **TagEditor** is an editor that works in combination with Workbench. Users can open one 'translation segment' at a time and either translate or edit the suggested translation extracted from the translation memory database.

The PE time was recorded by means of the standard feature of Trados combined with a Windows macro devised for this project to achieve thorough time measurements.¹⁶ The session was held on the 20th, 27th, and 28th of November, 2008, in Symantec's office in Dublin. The post-editors were provided with written instructions, which included a project summary, task instructions, and most importantly, PE guidelines. PE guidelines emphasised the quality requirements for post-edited text; the target text had to convey the correct meaning of the ST, and comply with Japanese grammar. However, stylistic modifications were discouraged, and the post-editors were instructed to not make changes to any user interface (UI) terms.¹⁷ The contents of the written instructions used for this preliminary study will not be discussed in detail, but the revised version used for the main study will be discussed in section 5.4.4.1, and the entire contents (written in Japanese) and its English translation will be presented in Appendix A.

The post-edited texts were compared with the MT output to calculate the amount of editing by means of the three AEMs mentioned in section 3.4.2. This preliminary study was approved by the DCU Research Ethics Committee.

4.3 Initial findings

4.3.1 Correlation between the amount of editing and PE effort

Table 4.1 shows the raw PE speed data: the mean value, standard deviation (SD), minimum value, and the maximum value, for post-editor A, B, and C. As can be seen from the table, the mean PE speed differs depending on the participants. In order to see the general trend across post-editors, these differences have been controlled for during the statistical analysis (section 4.4.5.3).

Post-Editor	Mean	SD	Min	Max
A	18.08	15.83	0.74	100
B	33.43	21.01	1.43	150
C	36.37	22.10	2.54	150

Table 4.1 Summary statistics for PE speed (words/min)

¹⁶ See section 5.1.1.2 for a detailed discussion

¹⁷ UI includes names for menus, dialog boxes, options, and so on.

Table 4.2 shows the results for AEM scores, representing the amount of editing (textual changes), calculated on a sentence-by-sentence basis from the data of all three post-editors, thus the number of observations (shown by ‘N=’) is 1,425 (475 sentences edited by three post-editors).

	Mean	SD	Min	Max
GTM	0.75	0.21	0	1
TER	28.31	33.06	0	300
BLEU	0.48	0.37	0	1

Table 4.2 Summary statistics for automatic metric scores (N=1,425)

Figures 4.2, 4.3, and 4.4 show a set of scatter plots from each post-editor’s results that depict the relationship between PE speed (y-axis) and AEM scores (x-axis).¹⁸ Each data point represents a sentence, and a Pearson correlation coefficient is shown in parentheses above each plot. Since the raw PE speed data in the unit of words per minute have a positively skewed distribution and have a slightly exponential relationship with automatic metric scores, we employed a method of transforming the data to logarithm numbers to ensure a normal distribution and a linear relationship with AEM scores to make sure they are suitable for statistical analysis (Woods et al. 1986).

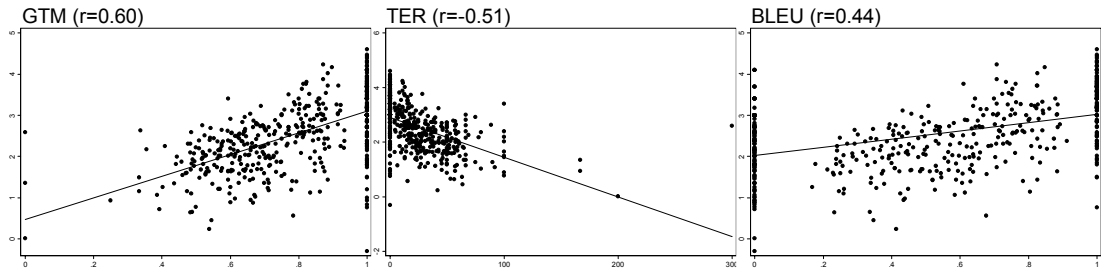


Figure 4.2 Correlation between AEM scores and logarithm of PE speed (Post-Editor A)

¹⁸ GTM allows to specify ‘exponent’ value to penalise incorrect word order. While it has been reported that the smaller exponent results in a better correlation with human evaluation in terms of adequacy, and the larger exponent results in a better correlation with human evaluation in terms of fluency (Lin & Och 2004), as a result of testing with different settings in this preliminary study, it was found that exponent 1.2, which mildly penalises the word order difference (Callison-Burch et al. 2007), had the highest correlation with PE speed. The same exponent setting will be used for the main study.

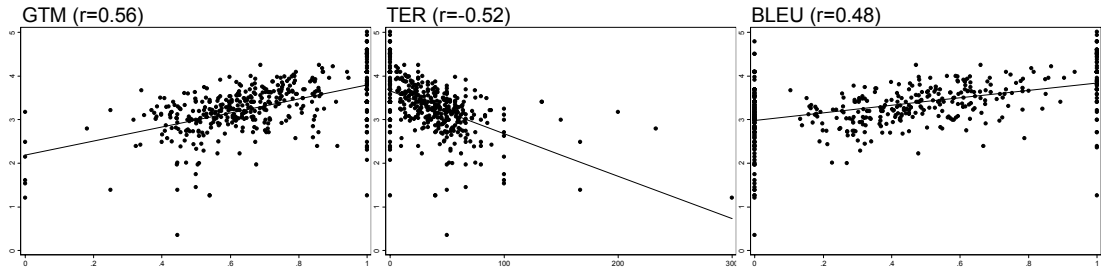


Figure 4.3 Correlation between AEM scores and logarithm of PE speed (Post-Editor B)

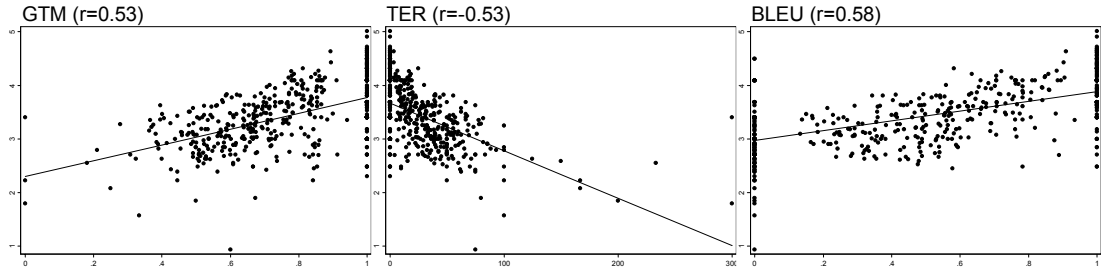


Figure 4.4 Correlation between AEM scores and logarithm of PE speed (Post-Editor C)

The shapes of the distribution are metric-dependent, with TER having a slope in the opposite direction as it gives a zero score for a perfect match between MT and PE output and increases the score as the distance between the two becomes larger. The groups of dots on the right edges of GTM and BLEU and the left edge of TER represents the ‘perfect match’ between MT and PE. The main reason that BLEU also has the groups of dots on the left edge is that its standard 4-gram option gives a zero score to all sentences that are shorter than four words, even if they are a perfect match. While GTM and BLEU accommodate all possible distances within the range between 0 and 1, TER does not have an upper limit (the distance between MT and PE depends partly on the difference in lengths of the two texts), and thus tends to produce outliers.

Despite these differences among the AEMs presented here, these figures altogether suggest that the relationship between the amount of editing (as measured by AEM scores) and the amount of PE effort (as measured by speed) is somewhat proportional. However, the distribution of the data points is still broad, which means that the relationship contains rather large variance, which may be explained by taking into consideration other factors. For the purpose of this preliminary study, we aimed to take into account a small set of variables related to ST characteristics and PE operation types.

4.4 Effect of combined variables on PE effort

Operationalisation of the concepts of ST characteristics and PE operation types have been discussed in section 3.4, in which we mentioned that actual variables need to be identified from the existing literature or in an exploratory manner. The following sections explain how such variables were discovered.

4.4.1 ST structure

It is reasonable to assume that the sentences with more complex structures require more effort to understand, translate, and edit for humans. Leech (2006) suggests three major categories of English sentence structures: a *simple sentence* contains only one clause, a *compound sentence* contains two or more clauses linked by coordination, such as ‘and’ and ‘but’, while a *complex sentence* contains one or more subordinate clause(s). Examples of each category taken from the test corpus are shown below.

Simple sentence:

- *Delete the item from the vault.*
- *An envelope with a paperclip indicates an email with one or more attachments.*

Compound sentence:

- *The shortcut is a direct link to the archived item, and it has the following icon.*

Complex sentence:

- *Select the items that XXX is processing.*
- *Put the item in the Restored Items folder in the mailbox that is specified in the Settings dialog box.*

While analysing all the source sentences and manually classifying each sentence into one of these categories based on Leech’s definitions, we became aware of a need for setting up another category: *incomplete sentence*, in other words, textual fragments consisting of words and phrases. Technical documents tend to contain a number of sentences or ‘segments’ that lack the basic structure of a sentence, that is, subject and

predicate, and cannot stand alone as complete sentences, and thus do not fall into any of the aforementioned sentence categories, for example:

Incomplete sentence:

- *File size*
- *For a file system vault:*
- *If there is more than one page of search results:*

These often appear as titles, headings, and cell contents in tables. In fact, 200 sentences out of 475 analysed in this preliminary study were incomplete sentences. In order to accommodate all types of sentences, we define in our study ‘a sentence’ as any type of complete or incomplete sentence. We also employ the term ‘segment’ to cover all types of complete and incomplete sentences, especially when we refer to translation units as recognised by TM and MT systems. This term also helps to keep in line with other studies in the same field (O’Brien 2006b, O’Brien 2007, Guerberof 2008). The word ‘sentence’ and ‘segment’, therefore, will be used interchangeably throughout this study. It also came to light that there were very few compound sentences in the test corpus; only four such occurrences were found. The reason for this is not clear, as compound sentences are not explicitly banned in Symantec’s controlled language rules. However, it may be the sentence length limitation of 25 words that encourages the splitting up of one compound sentence into two separate sentences. In any case, since we have too few observations of compound sentences, we excluded this category from further analysis in the preliminary study.

As shown in Table 4.3, a comparison of the Pearson correlation coefficients of amount of editing, measured by GTM, and PE effort, measured by speed, between the sentence structures suggested that simple sentences have stronger correlation compared to other sentence types. From this observation, we suspected that sentence structure has some effect on the amount of PE effort.

	Simple (N=143)	Complex (N=128)	Incomplete (N=200)
Post-Editor A	0.73	0.60	0.51
Post-Editor B	0.72	0.54	0.56
Post-Editor C	0.73	0.66	0.53
Average	0.73	0.60	0.53

Table 4.3 Difference in correlation levels between sentence structures

4.4.2 ST length

The sentence length often becomes an issue for writing the ST for MT, and thus CL rules and writing style guides often include a rule that limits sentence length, for example, 20-25 words maximum (Kohl 2008). Very long sentences can entail both grammatical and semantic complexity, which is problematic for both humans and MT. On the other hand, extremely short sentences often lack context and thus may be semantically ambiguous for both humans and MT. O'Brien (2006b) reports that short sentences tend to cause more PE effort in the English-German language pair.

We need to pose the question then, which is easier to post-edit, longer sentences or shorter sentences? In an effort to answer this question, we checked the average PE speed of each post-editor by ST length dividing sentences into categories according to the number of words: 1-5, 6-10, 11-15, 16-20, 21-25, and over 25 words. The results are shown in Table 4.4 and Figure 4.5.

ST length in words:	1-5 (N=179)	6-10 (N=91)	11-15 (N=86)	16-20 (N=68)	21-25 (N=40)	26 - (N=11)
Post-Editor A	19.2	23.7	16.9	14.0	12.4	7.9
Post-Editor B	35.9	39.6	30.4	27.2	27.7	24.6
Post-Editor C	32.9	45.0	37.9	33.8	35.5	28.0

Table 4.4 Average PE speed (words/min) for different ST length groups

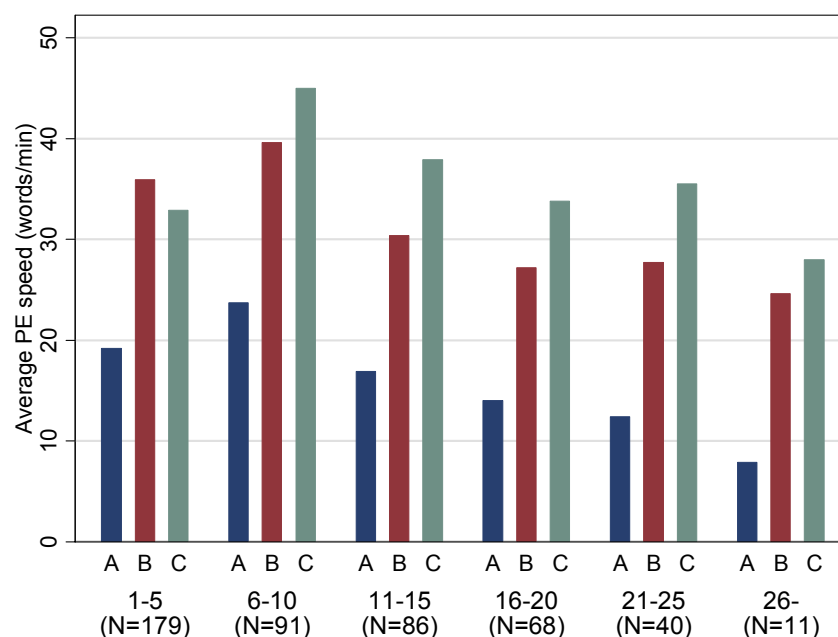


Figure 4.5 Average PE speed by ST length and post-editors

As can be seen, the sentences of six to ten-words long are faster to post-edit than all other ST lengths for all three post-editors. According to this result, the PE speed is rather slow for sentences containing five words or fewer, becomes faster as the sentences become longer, and from a certain point on the PE speed becomes slow again as the sentences become longer. This finding led to the decision to include the source sentence length as one of the variables that affects the amount of PE effort.

4.4.3 Observed common PE operations

Similar to the English function words and content words, morphs in Japanese can also be categorised into either *function morphs* or *content morphs*. According to Leech (2006), function words (also known as ‘grammatical words’) are defined by their grammatical function, such as determiners, prepositions, conjunctions, auxiliary verbs, and pronouns, as opposed to content (or ‘lexical’) words, such as nouns and verbs. Groves & Schmidtke, identifying the PE operation patterns for English to German and English to French MT output of Microsoft documents, report that editing of function words occurs far more often than editing of content words, with determiner changes occupying 70% for French and 42% for German (Groves & Schmidtke 2009). In our preliminary study, a similar trend for function/content morph ratio was observed; 64% of all edits were performed on functions morphs, 49% of which are on various *particles*.

Japanese particles are generally divided into four groups: case particles, conjunction particles, adverbial particles, and sentence-final particles (Hayashi et al. 2004). *Case particles* are usually attached to substantives¹⁹ and show the relationship between the preceding and succeeding morphs. For example: (particles are shown with double underlines)

My father drove my car and went to Tokyo with my mother.

父 が わたし の 車 を 運転して 母 と 東京 へ 行った

Subject indicator Possessive indicator Object indicator "with" "to"

Conjunction particles, on the other hand, show the conjunctive relationship between morphs before and after the particle, and are attached to predicates. For example,

That is over my budget, but I will buy it if the quality is good.

予算を超える が 品質が良けれ ば 買おう。

"but" "if"

Adverbial particles are attached to substantives and other particles and modify them in the same way as adverbs. For example,

He survived with only water for about a week.

彼は一週間 ほど 水 だけ で生き延びた。

Subject indicator "about" "only"

Sentence-final particles can be attached to various types of words, but are usually placed at the end of the sentence to add meanings such as question, prohibition, and desire. For example,

¹⁹ In Japanese part of speech categories, substantives, or *taigen*, are nouns or words that act as nouns, which do not inflect.

What is this?

これはなんですか。
└
Interrogative
indicator

Among all the edits involving particles, 75% involved case particles, and 10% conjunction particles, which altogether accounts for about 27% of all the edits observed in this preliminary study. As can be seen from the examples above, a change in a case particle or a conjunction particle often involves the change in the relationship between words and phrases, leading to a change in the meaning of the whole sentence, which we suspect is an effort-intensive PE task. The relationship between words is called ‘dependency’.

4.4.4 Dependency

In Japanese sentence analysis, dependency is regarded as one of the basic processes, and has been studied by a number of researchers (Uchimoto et al. 1998, Kudo & Matsumoto 2002, Abekawa & Okumura 2006, Iwatate et al. 2008). Abekawa’s remark about ‘dependency’ concisely defines the term when used for the Japanese language (ibid: p. 833).

The Japanese dependency structure is usually represented by the relationship between phrasal units called bunsetsu, each of which consists of one or more content words [morphs] that may be followed by any number of function words [morphs].

Based on the classified edit patterns, which will be explained in section 6.2.1, the author identified the edits that involved dependency changes, and marked the sentences that included one or more dependency edits to distinguish them from other sentences that did not require any dependency editing. An example of such dependency change is shown below (Glosses are shown in brackets). In this case, the English preposition was incorrectly interpreted by MT to the Japanese case particle ‘*の*’ instead of ‘*で*’, which consequently changed the relationship between ‘items’ and ‘vaults’, and also misinterpreted what ‘one or more’ modifies.

English ST: Search for items in one or more vaults.

MT output: 1 つまたは複数の Vault の アイテムの有無を検索します。 [Searches for one or more items of vault.]

PE result: 1 つまたは複数の Vault で アイテムの有無を検索します。 [Searches for items in one or more vaults.]

It was found that the PE speed is slowed down when dependency editing was necessary; the average PE speed of all three post-editors for the sentences that required dependency editing was 20.2 words/min, while for the sentences that did not require dependency editing it was 32.2 words/min. This suggests that dependency editing systematically increases the PE effort.

4.4.5 Multiple linear regression analysis

After confirming the existence of the systematic effect of these variables, namely, ST structure, ST length, and dependency changes, we took one step further and attempted to examine if each variable has an effect independent of other conditions. This was made possible by MLR discussed in section 3.5. In sections 4.4.1, 4.4.2, and 4.4.4, we examined the effect of each variable independently, but a question still remains as to whether each of them holds the effect when other conditions are also taken into account, for example, if dependency editing still slows down the PE process if we compare sentences with the same structure and length. PE effort may be more effectively analysed by taking into account a number of possible conditions at the same time.

For simplicity, we decided to use only GTM as the indicator of the amount of editing from among the three AEMs, since it had the highest average correlation with PE speed, and it did not tend to produce outliers, as mentioned in section 4.3.1.

4.4.5.1 Use of ‘square term’ for sentence length

In order to examine if both very short and long sentences slow down the PE speed as indicated in section 4.4.2, we used two variables: the ST length, and the *square* of the ST length, which is called ‘quadratic functions.’ Quadratic functions are often employed to capture diminishing or increasing effect of an independent variable on the dependent variable (Wooldridge 2006). In practice, the square of an independent variable is

introduced along with the normal term, the non-squared original variable. If they are assigned coefficients in opposite signs as a result of MLR, we know that the independent variable has a non-linear effect, that is, if the normal term is positive and the square term is negative, it has a diminishing effect (Figure 4.6, left), if the normal term is negative and the square term is positive, it has an increasing effect (Figure 4.6, right).

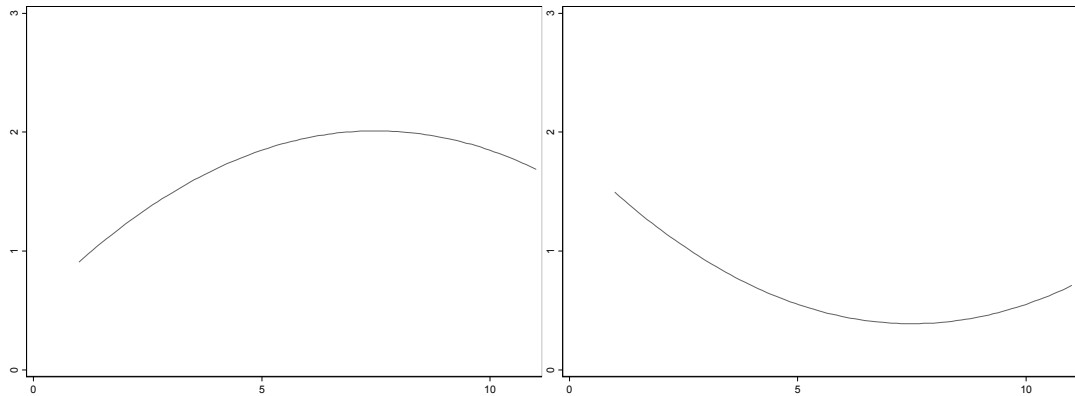


Figure 4.6 Graphical representation of quadratic function

4.4.5.2 Binary variable for dependency edits

In order to see the effect of dependency editing on the amount of PE effort, we employed an independent variable to indicate the presence or absence of dependency editing. For the purpose of this preliminary study, we decided to use a binary variable for simplicity: 0 for sentences with no dependency edit, and 1 for sentences with one or more dependency edit(s).

4.4.5.3 Results and discussion

Table 4.5 shows the results of the multiple regression analysis. As different sentence structures seemed to have different distribution in the correlation between the amount of editing and the amount of PE effort, as discussed in 4.4.1, we ran the regression analyses for all sentences first, and then separately for sentences in each sentence structure category, namely, Simple, Complex, and Incomplete sentences. Also, in order to take into account the difference in average PE speed between participants, additional independent variables, Post-Editor A and Post-Editor B, were employed to cancel out

the difference; the coefficient for Post-Editor A and Post-Editor B represent the speed difference in relation to Post-Editor C.

Model:	I	II	III	IV
Sample:	All sentences	Simple sentences	Complex sentences	Incomplete sentences
Independent variables:				
GTM Score (Range: 0 - 1)	2.208*** (0.08)	2.720*** (0.15)	1.863*** (0.17)	1.811*** (0.12)
ST Length (Range: 1 - 41)	0.070*** (0.01)	0.071*** (0.02)	-0.000 (0.02)	0.225*** (0.02)
ST Length^2 (Range: 1 - 41)	-0.002*** (0.00)	-0.002*** (0.00)	0.000 (0.00)	-0.011*** (0.00)
Dependency Edit (0: Absent / 1: Present)	-0.224*** (0.04)	-0.022 (0.05)	-0.245*** (0.05)	-0.394*** (0.07)
Post-Editor A	-0.926*** (0.04)	-0.927*** (0.05)	-1.255*** (0.05)	-0.714*** (0.06)
Post-Editor B	0.018 (0.04)	-0.021 (0.05)	-0.035 (0.05)	0.102 (0.06)
Adjusted R-squared	0.54	0.63	0.69	0.50
Number of cases	1,413	429	384	600
Dependent variable: Logarithm of post-editing speed (Range: -.301 to 5.011, Standard Deviation: 0.799) Unstandardised coefficients are shown in bold face, with standard errors in the parenthesis underneath. Asterisks indicate statistical significance. ***: p < 0.001				

Table 4.5 Regression analysis results of PE speed by ST structures

The effect of each variable on increasing PE speed (decreasing PE effort) is shown as coefficients as a result of one unit increase of each variable. Statistical significance is shown by asterisks; three asterisks mean that the p-value is less than 0.001. Model I shows the results from the entire sample, while II, III, and IV show the separate results by sentence structures.

The values for GTM Scores show how much GTM scores, holding other conditions fixed, correlate with PE speed after taking into account other conditions listed in the table. The PE speed for GTM=1 sentences (perfect match), is approximately 14 times, 5.5 times, and 5 times faster compared to GTM=0 sentences for simple, complex, and incomplete sentences respectively.²⁰ In order to see the effect in a more granular manner, we broke it down into a unit of 0.1 and recalculated the coefficients, which suggested

²⁰ The coefficient calculated from a logarithmic number can be converted to a multiplier by the equation: $\exp(\text{coefficient}) - 1$. In this case, $\exp(2.720) - 1 = 14.2$, $\exp(1.863) - 1 = 5.4$, and $\exp(1.811) - 1 = 5.1$ respectively.

that the increase of 0.1 in the GTM score increases the PE speed by approximately 31%, 21%, and 20% for simple, complex, and incomplete sentences respectively.²¹ In either unit, it can be seen that GTM scores have a stronger relationship with PE speed for simple sentences compared to complex and incomplete sentences, which confirms that our earlier findings in section 4.4.1 hold even after controlling for other variables.

The coefficient for ST Length for incomplete sentences is positive (0.225) and its squared term is negative (-0.011). This is evidence for the diminishing effect of ST length on PE speed, as explained in section 4.4.5.1. According to the estimate shown in the table, the threshold is 10/11²²; for sentences shorter than 11 words, holding other conditions fixed, the PE speed becomes faster as the sentence length approaches 10 words, and for sentences longer than 10 words, the PE speed becomes slower as the sentence length becomes longer. A similar effect can be observed for simple sentences, but the evidence is weaker (0.071 and -0.002). The estimate shows that the threshold for simple sentences is 15/16.²³ As for complex sentences, the statistical significance for this variable was not obtained. One of the possible explanations for these diminishing effects may be that the longer sentence helps disambiguation up to a certain point, but also increases semantic and/or grammatical complexity from that point onward.

Dependency also seems to have different effects on different sentence structures. According to the estimate, presence of one or more dependency edit(s) slows down the PE speed by 32.6% and 21.7% in incomplete and complex sentences respectively, compared with only 2.2% in simple sentences.²⁴ The result for simple sentences is not statistically significant either. In the case of complex sentences, it is intuitively understandable that dependency edits in sentences, whose structure is complex, may result in a much more effort-intensive task than sentences with a simpler structure. In the case of incomplete sentences, on the other hand, the inherent contextual ambiguity may be one of the causes for a high level of effort.

²¹ Similarly, $100 * (\exp(.2720) - 1) = 31.3$, $100 * (\exp(.1863) - 1) = 20.5$, $100 * (\exp(.1811) - 1) = 19.9$.

²² $(0.2251739) * \text{Sentence Length} + (-0.0114461) * \text{Sentence Length}^2$ becomes larger towards Sentence Length=10, and smaller after Sentence Length=11.

²³ $(0.0706634) * \text{Sentence Length} + (-0.0023444) * \text{Sentence Length}^2$ becomes larger towards Sentence Length=15, and smaller after Sentence Length=16.

²⁴ $100 * (\exp(-0.394) - 1) = -32.6$, $100 * (\exp(-0.245) - 1) = -21.7$, $100 * (\exp(-0.022) - 1) = -2.2$

The variables for Post-Editor A and Post-Editor B show the overall speed differences in comparison to Post-Editor C. These variables have been introduced only for the purpose of cancelling out the between-subject differences here, and further analysis about the differences between post-editors will be postponed to Chapter 8.

4.5 Lessons learnt

These findings fulfilled the first purpose of the preliminary study: to obtain tentative answers to supplement the operationalisation process. Through the process of qualitative analysis, combined with a literature review, we identified some factors that affect the amount of PE effort, such as sentence structure, sentence length, and dependency editing, which will be used as operationalised variables in the main study.

This preliminary study also served the second purpose. We had an opportunity to test our proposed method and tools. It proved that the tools did not have any major technical flaws, but identified a problem in the test corpus. The test corpus used for this study contained a number of exact duplicates of sentences, which was found to have caused problems in keeping precise time records during PE. Therefore, it was decided to remove all the duplicate sentences in the test corpus for the main study (see 5.2.2.1 for a detailed discussion on this issue).

In addition, the preliminary study helped to select the most suitable AEMs for use in the main study. GTM, the one that had the highest correlation with PE speed will be used in the main study, along with TER, which provides detailed editing information in a plain text format and which helps to classify PE operations in a consistent manner. The refined methods are fully explained in Chapter 5.

Another interesting finding was that, despite the fact that it was clearly stated in the post-editing guidelines that UI terms should not be changed, post-editors A and B have made changes to some of the UI terms: 12 and 2 out of 186 UI terms, respectively. This issue will be discussed in more detail in section 5.2.3.1, and further investigated in the main study.

4.6 Concluding Remarks

This chapter outlined the purposes, methods, and findings of the preliminary study. Based on the initial findings, it proceeded to define other variables by combining qualitative and quantitative analysis. It then examined a combined effect of all variables considered by means of a statistical analysis technique, MLR, which helped us to understand complex relationships between various conditions and the amount of PE effort. Finally, it reviewed the process and the findings of the preliminary study, in which it was confirmed that the purpose of the preliminary study were fulfilled and some insightful input was obtained to refine the methods for the main study. The methods and the results of this preliminary study have also been presented at and published by MT Summit XII (Tatsumi 2009).

Chapter 5 METHODS OF DATA COLLECTION AND ANALYSIS

This chapter details the practical methods employed in this study, which are grounded on the research methodology and the outcome of the preliminary study discussed in Chapters 3 and 4 respectively. This chapter is organised in such a way that each section focuses on different parts and phases in the present study. We first discuss how the operationalised definitions were measured in practice in section 5.1. Then, a series of practical methods employed to carry on the experiment will be explained in sections 5.2, 5.3, 5.4, and 5.5.

5.1 Measurement Technique

The conceptual definitions used in the research questions have been partly operationalised in Chapters 3 and 4. The rest of the variables are to be discovered in the Exploratory phase of the main study (Chapter 6). This section explains how some of the operational definitions were measured in practice.

In the process of operationalisation, we considered mainly *measurement validity* in order to ensure that the operational definitions reflect the conceptual definitions we aim to measure. Here in discussing what techniques to use, we focus on *measurement reliability* in order to ensure that the chosen techniques are capable of measuring the concepts in a consistent and stable manner. Any measurement contains two types of errors (gaps with the true values), namely, *random errors* and *systematic errors* (Boslaugh & Watters 2008). Random errors occur by chance and are unpredictable, as human beings or even measurement instruments may not perform exactly the same way in repeated measurements. They are therefore unavoidable. However, since random errors are considered to cancel each other out and the average value of random errors in a large number of measurements is assumed to be zero (Boslaugh & Watters *ibid*), we do not have to worry about it too much. The more serious problem is the systematic error, which is caused by a biased measurement method, and has a consistent pattern. The measurement method is under the control of the researcher (Frey et al. 1991), therefore although it is impossible to eliminate all the systematic errors, researchers

should strive to reduce them as much as possible, or be aware of the errors when analysing the results.

According to Frey et al. (1991), measurement reliability is generally higher with quantitative measures because of their more formally structured nature than qualitative measures. Although the main measures of the present study are quantitative, in order to avoid human errors and further increase measurement reliability, we employ mechanical methods as often as possible. The following is a detailed explanation of the measurement technique employed for the variables in the present study.

5.1.1 PE speed

The amount of PE effort was operationalised in section 3.4.1 into a measurable variable, PE speed. Here we explain how to actually calculate the speed, which is the word count of an ST sentence (English) divided by the number of minutes taken to post-edit the sentence.

5.1.1.1 Word count

A number of text processing software packages, including MS Word and SDL Trados Translator's Workbench, offer word counting functions, but it is preferable to use a single software program to count the word number for all sentence by sentence analyses in order to ensure consistent measurement. In this study, since Microsoft Excel was used to organise all the sentence data, the same software was used to count the word number by using the following equation:

$$=IF(LEN(TRIM(A1))=0,0,LEN(TRIM(A1))-LEN(SUBSTITUTE(A1," ",""))+1)$$

where A1 is the name of the cell that contains the sentence whose word number needs to be counted.

5.1.1.2 Time

The time data were recorded primarily by means of a default function of SDL Trados Translator's Workbench version 7, which automatically records the year, month, day,

hour, minute, and second of the time when the translation sentence has last been saved. However, relying on this function is somewhat problematic since if the same segment is opened and closed more than once, the program overwrites the saved date and time after the second time, which means that the previously recorded time is lost. In order to overcome this problem, we devised a macro using a freeware key macro program called AutoHotKey. This macro is designed to force Trados Translator's Workbench to freshly save the segments every time they are opened and closed. It is activated when post-editors press the standard Trados TagEditor key sequence for closing the segment, that is, Alt-End. However, even using this solution, another problem of using the Trados time stamp function remains. As it does not record the time when the segment is opened, we need to assume that the time when the post-editor finished editing and closed the previous segment is the time the post-editor opened and started editing the current segment. This without doubt does not reflect the exact 'pure' editing time as it may contain a delay of a few seconds for sipping coffee, stretching shoulders and so on. Although Trados TagEditor provides an option that saves and closes the segment and opens the next segment immediately, this does not solve the fundamental problem as it cannot exclude such time delay either. This is not a random error, but a systematic error, as the gap with a true value is always a delay, thus the average of this error will not be zero, which may compromise the measurement reliability. Being aware of this issue, we still decided to use the Trados time stamp function for a number of reasons. First, using one of the standard CAT tools instead of devising a special text editing facility for this study enables us to conduct the study using software familiar to post-editors, thereby increasing ecological validity. Secondly, Trados Translator's Workbench provides the time data in numbers in a simple text format, which makes the large amount of data manageable to process. Finally, using a single software program for both PE and time keeping makes the experiment procedure simple, rather than using two different software programs and having to synchronise them. However, for the purpose of triangulation, we will also consult the screen capture data (discussed in 5.1.5) when the time data from Trados Translator's Workbench needs qualitative information to clarify why certain edits have taken post-editors excessive time.

The word number of a sentence calculated using MS Excel was then divided by the number of minutes (including the number of seconds converted to a fraction of a

minute) taken to edit the sentence recorded by Trados Translator's Workbench to obtain PE speed data.

5.1.1.3 Accuracy of speed data

Threats to measurement reliability still exist. One of the possibilities is that the post-editors may gain speed over the course of the entire session because of an increase in task familiarity, or may lose speed due to boredom or fatigue. Therefore, after the data collection and by means of a t-test, it was confirmed that there was no such systematic effect. Another possibility is that the post-editors may fail to perform certain operations required to keep the precise time record, such as pressing the specified keys to activate the key macro. In an effort to avoid this problem, the importance of required operations was emphasised in both written and oral instructions prior to the experiment, as will be discussed in section 5.4.5. Participant post-editors were also requested to report any failure to follow the precise instructions.

5.1.2 Automatic Evaluation Metrics (AEM)

The use of AEM as a variable to measure the amount of editing has already been discussed in section 3.4.2. We chose to use GTM for the purpose of quantifying the technical effort, guided by the findings from the preliminary study discussed in Chapter 4. In addition, we employed TER for the purpose of qualifying the ingredients of editing, which will be discussed in Chapter 6. As the Japanese writing system does not insert spaces to mark the boundary of words, the text was separated into morphs by means of a tokenising program.

5.1.2.1 Tokeniser

Japanese text is written without spaces to separate words; in fact, the definition of a 'word' in the Japanese grammar differs depending on different schools of grammatical theory (Iwabuchi 2000, Hayashi et al. 2004). The more common concept in the field of Japanese natural language processing is *morph*. The smallest unit of meaning is defined as a 'morpheme' (Yoshimura 2000), and 'morph' is the actual linguistic form of a morpheme. To give an example in English, the word 'unforgiving' consists of three morphs: 'un', 'forgiv(e)', and 'ing', each of which is recognised as a different morpheme: a boundary morpheme, a free morpheme, an inflectional morpheme,

respectively. Determining the boundary of a morph in Japanese, however, is not straightforward since multiple interpretations are possible depending on where the text is separated. To resolve such ambiguity and segment the text properly, Japanese tokenisers normally perform morphological analysis, inflection analysis, and part of speech tagging using specialised dictionaries. As different programs perform these analyses differently, the tokenisation results also differ from each other, thus there was a need to find a suitable one for the present study. A number of Japanese tokenisers are available as either open licence products, such as ChaSen,²⁵ MeCab,²⁶ Juman,²⁷ and TinySegmenter,²⁸ or commercial products, such as Rosette,²⁹ Gengoro,³⁰ and Marimo.³¹ Because of limitations on budget, only open licence products were considered in the present study, and based on the testing and comparison, MeCab was chosen for the ease of use and the relatively smaller number of segmentation errors observed in our corpus.

5.1.3 Characteristics of source text

We considered two variables in terms of characteristics of ST in the preliminary study.

The *ST length* is counted by the ST word number, which can be done by means of an Excel function as discussed earlier in 5.1.1.1, and this ensures measurement reliability.

The *ST structure* includes simple sentence, compound sentence, complex sentence, and incomplete sentence, as discussed in section 4.4.1. To the best knowledge of the researcher, there is no software program that automatically detects the difference in sentence structure, thus it needs to be classified manually by the researcher, which could

²⁵ Developed by Nara Institute of Science and Technology. Accessible from: <http://sourceforge.jp/projects/chasen-legacy/> [Last accessed: 19/2/2010]

²⁶ Developed by Kyoto University and NTT. Accessible from: <http://mecab.sourceforge.net/> [Last accessed: 19/2/2010]

²⁷ Developed by Kyoto University. Accessible from: <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html> [Last accessed: 19/2/2010]

²⁸ Developed by Kudo Taku. Accessible from: <http://chasen.org/~taku/software/TinySegmenter/> [Last accessed: 19/2/2010]

²⁹ A product of Basis Technology Corporation. Information available from: <http://www.basistech.com/base-linguistics/asian/> [Last accessed: 19/2/2010]

³⁰ A product of ZOO Corporation. Information (in Japanese) available from: <http://gengoro.zoo.co.jp/> [Last accessed: 19/2/2010]

³¹ A product of Mooter. Information (in Japanese) available from: <http://enterprise.mooter.co.jp/marimo/> [Last accessed: 19/2/2010]

weaken measurement reliability if not done according to unequivocally stated definitions. Therefore, we decided to follow Leech's definitions (Leech 2006) for simple, compound, and complex sentences, and classified any textual fragments consisting of words and phrases as incomplete sentences, as mentioned in section 4.4.1.

5.1.4 Types of PE operations

Types of PE operations were identified based on a qualitative investigation of obtained data. Detailed discussion on discovering PE operation types will be discussed in section 6.2.1.

5.1.5 Observation of PE operation

Observation of PE operation was carried out by means of a screen capture software program. Screen capture software programs record on-screen activities into a video file that can be viewed on any computer to see what operations have occurred on the computer. One can replay the screen activities and see how the computer operator moved the cursor and the mouse, typed, inserted, deleted, copied, and pasted text, chose menus and options, and so on, by looking at the screen activities just as the operator saw her/his own screen. Such software programs do not use external cameras, nor record post-editors themselves. Therefore, the observation of PE activity is limited to their actions on the computer, and any body movements, facial expressions, or utterances are not taken into account in this study. This is appropriate since we want to focus on the way post-editors conduct their professional work as manifested on the computer, and psychological analysis is beyond our scope. Unlike videotaping with a standard camera, in which case different settings, such as lighting, angle, and zoom may result in inconsistent recording of data, screen recording software can work in a consistent manner on every computer, thus making possible a standardised way of gathering data, which contributes to measurement reliability. In addition, using screen capture software helps to maintain participants' normal work environment compared to setting a video camera in their workplace or having an observer monitoring their behaviour. Screen capture programs have been increasingly employed in the field of behavioural research ((Arapakis et al. 2008, Loizides & Buchanan 2009, Tort et al. 2009), and also in translation and PE activity observation (Désilets et al. 2008, de Almeida & O'Brien

2010). Among available products, we used BB Flashback³², as its trial version was downloadable free of charge, and was found to be user-friendly and less CPU intensive compared to other programs.

5.1.6 Post-editors' attributes

The information about post-editors' attributes were gathered by means of a questionnaire. Out of 34 questions contained in the questionnaire, 20 were about their background and experience in translation, post-editing, and tools. The contents of the questionnaire will be discussed in detail in 5.5.

5.2 Test Corpus Design

In the experiment conducted for the present study, a parallel corpus of English ST and the Japanese MT output was used for post-editing by nine post-editors. This section describes the profile and the method of compiling the test corpus.

5.2.1 Sampling considerations

5.2.1.1 Brief profile

The text chosen for the present research was extracted from a user manual of a data storage product developed by Symantec corporation. This product was selected because it was one of the latest MT projects for which controlled language rules, user dictionaries, and pre- and post- processing tools were used in fullest force. TM match segments from previous versions of the same product were also available for this project, which made it possible to study the difference between post-editing of MT output and editing of TM match segments. The reason for the sole use of Symantec documentation is a) the researcher of the present study was granted access to their past documentation for the research, b) Systran, the MT system used for this research, makes extensive use of the customised dictionary, and the researcher was given full access to Symantec's dictionary in order to optimise the system; if documents from other sources were used, the MT system could not have performed optimally, and c) Symantec has been

³² A product of Blueberry software. Information available from:
<http://www.bbsoftware.co.uk/BBFlashBack/Home.aspx> [Last accessed: 19/10/2010]

researching and adopting various technologies and tools to refine the output of Systran, thus the researcher had access to the optimised use of an RBMT system, making the research context highly valid to real-life industrial contexts.

5.2.1.2 Size

There were several considerations as to the size of the test corpus for this study. It had to be large enough to accommodate different ST attributes, such as sentence length and structure, MT output quality, and so on. At the same time, it had to be short enough to allow participant post-editors to work on the text without becoming disengaged from the research project. Therefore, we aimed to build a test corpus that contained the text for approximately one-day's work by a post-editor. The speed of PE, however, differs depending on experience, skills, methods, and so on. McElhaney & Vasconcellos (1988) reported that after about a month of full-time practice, post-editors tended to start feeling comfortable with their work, and the average throughput was about 6,000 words of final quality text per day. More recently, Plitt & Masselot (2010) have conducted a PE experiment using a large amount of text, from which throughput of PE was found to be 800 to 1,800 words per hour depending on post-editors. TAUS has reported, from a survey result, that post-editors' daily throughput is about 3,800 – 5,600 words per day when publishing quality is required (TAUS 2010). In our preliminary study, a 4,784-words test corpus was used, which took experienced translators less than one day and an inexperienced translator one and half days to post-edit. Based on this, it was assumed that 5,000 words would be an appropriate amount to be post-edited in one day by experienced translators/post-editors. The master corpus (the entire manual) contained 55,349 words in 5,815 sentences in 195 files, from which the test corpus was compiled consisting of 5,029 words in 359 sentences in 32 files.

5.2.1.3 Format

The original documentation was written in XML format, which was used as is in the test corpus. This was appropriate because XML tagged format is widely used not only in Symantec but also in IT documentation in general. More importantly, as this study was designed to be conducted in a condition as close as possible to the real-life work environment, it is of interest to observe PE activities performed on text written in a popular format to see how tags may affect PE.

5.2.1.4 Document component parts and text types

Byrne (2006) maintains that the roles of software user documentation are to define concepts, explain procedures, and provide the sources of additional information in order to familiarise the audience with the software and help them to use it. He also points out that software user documentation usually consists of independent ‘modules’ that contain a set of information about completing a task. The corpus employed in the present study meets these definitions in that each file serves as a section or a subsection of a user manual containing a title, a brief description, a detailed explanation of either concepts or procedures, and a list of additional information resources.

Also, Byrne (ibid: p.50), introducing four categories for technical documentation, namely, ‘Procedural documents’, ‘Descriptive and explanatory documents’, ‘Persuasive or evaluative documents’, and ‘Investigative document’, states that the first two categories relate directly to software documentation. Some other researchers also distinguish descriptive and procedural (instructive) passages in IT-related documentation (O’Brien 2006b, Roturier 2006). A *procedural* passage lists the steps needed to perform certain actions, such as ‘Click the target or target group that you want to edit’, while a *descriptive* passage explains a concept, such as ‘XXX keeps a master collection of all items that have ever been accepted into any case.’ There are also other types of text, such as titles and list items. In the present study, text types were identified by the XML tags embedded in the original document, and the effect of different document component types on the amount of PE speed was examined separately, which will be discussed in section 7.4.3.

5.2.1.5 Language pair

This study was conducted on the English to Japanese language pair. For IT documentation, the source language is most often English, and it is translated into various languages,³³ thus it is appropriate to choose English as the source language. The choice of the target language is due to several reasons: a) English to Japanese translation is high in demand in terms of volume and also strategically important (DePalma & Kelly 2009, Japan Translation Federation 2009), and thus researching this language

³³ For example, among all 28 job postings for IT-related translation in www.proz.com, 20 were from English to various languages. [Accessed on: 10/8/2010]

combination is of importance; however, b) little work has been published on post-editing of English-Japanese translation; and c) the researcher's ability to analyse the results limits the selection to Japanese. Although this study was conducted only in this language pair, the basic concept and approach may be applicable to other language pairs.

5.2.1.6 TM match segments

Although the main interest of the present research is the editing process for machine-translated text, we felt it was also interesting and relevant to compare this with the editing process of TM match segments, as combining TM and MT is becoming common in real industrial settings. For this reason, it was decided to include a small proportion of TM match segments extracted from the translation project of the previous version of the same product with a variety of match ratios. However, it is not realistic to edit TM segments with a very low match ratio. In fact, the appropriate threshold of effective TM match ratio has been one of the important topics of discussion since the introduction of TM technology. Bowker (2002) mentions the debate between clients and translators over appropriate TM match thresholds, where some clients want to set the threshold at 50% while translators find editing 50% match segments is more time consuming than freshly translating them, and the threshold is often settled at around 60-70%. Austerlühl (2001: p.141), on the other hand, shows an example price list for fuzzy match translation, in which 84% and under are regarded as new translation and charged at the full price. In the present study, it was decided to include TM segments with match ratios from 75% to 99%, to conform to Symantec's business practice. For all the segments under 75% match, along with the ones of 100% match, the translation in the TM was discarded. The reason for discarding 100% matches was because including 100% match segments would not give us any additional information on PE. Moreover, in practice, post-editors and translators are often told to not make changes to 100% match segments. The procedures of TM match segment selection is discussed in more detail in 5.2.2.2.

5.2.2 Procedures

Ideally, the test corpus should contain an adequate representation of various characteristics, including ST characteristics, such as ST length, ST structures, grammatical constructions, lexical elements, and so on, as well as various TM match

ratios. It would be impractical, however, to deliberately include all these elements in such a way that the test corpus is perfectly representative of the master corpus while keeping the natural flow of the text as a user manual. In this study, therefore, we make an assumption that the elements that are not included in a 5,000-word corpus are not frequently occurring elements in the master corpus of 55,000 words either, and therefore can be omitted from the investigation of the present study.

The first step to create the test corpus was selecting files from the whole master corpus written in English (section 5.2.2.1). The reason for not selecting text on a sentence-by-sentence basis from the beginning was, as mentioned earlier, to produce a test set that can be recognised by post-editors as a normal manual rather than an artificially constructed text of random sentences. The second step was to extract fuzzy match segments from the TM of a previous translation project (section 5.2.2.2). The third step was to machine-translate the rest of the text using Systran (section 5.2.3). The final step was to control the appearance of product-specific terms (section 5.2.3.1).

5.2.2.1 Selecting English source files

The main criterion in choosing appropriate files was that the files do not contain repetitions. Software manuals often contain duplicated sentences, but including such sentences in the test corpus, similar to including 100% match TM segments, would not give us any additional information as to PE tasks, since duplicated sentences need not be edited. More importantly, repetitive sentences cause problems in keeping accurate time records when using our current tools, which cannot keep a separate record for the translation of the identical source segment. Therefore, it was preferable to compile a corpus only from files with no repetitive sentences. There were 25 such files among all the files in the master corpus, but they consisted only of 4,180 words in 313 sentences altogether. In order to fill the shortage of the planned word count mentioned in section 5.2.1.2, seven other files with one or two repetitions were selected and repetitive sentences were eliminated in such a way that it would not compromise the natural flow of the text.

The exclusion of files that contained repetitions resulted in ruling out certain types of files, namely, files that consisted mainly of tables, as table headings and cell contents

tend to be reused more than once. This led to the test corpus having a smaller number of short and incomplete sentences compared to the master corpus. This contributed to having sentences with different structures and length in a more balanced way than the text used in the preliminary study (Chapter 4), where random sampling was employed, which resulted in having 200 incomplete sentences out of the total of 475 sentences. Although translation of short phrases and sentences can be problematic in automated translation and can be an important topic in its own right, we aimed at examining sentences with a broad range of lengths.

5.2.2.2 Extraction of TM segments

The composition of TM match segments was similar between the master corpus and the initial test corpus. However, to secure a sufficient amount of text for the analysis of post-editing of MT, some of the TM match segments were discarded and the ST was newly translated by MT, which resulted in having a smaller proportion of TM match segments in the final test corpus compared to the master corpus (see Table 5.1). All segments except for those whose translation were extracted from the previous TM were translated by Systran as will be discussed in section 5.2.3. Table 5.1 summarises the composition of TM match ratios in the master corpus, initial test corpus, and the final test corpus. The statistics were produced by the Trados Translator's Workbench 'Analyse' function³⁴, and the figures are based on the word count, and not the number of segments.

	Master corpus ³⁵	Initial test corpus	Final test corpus ³⁶
95-99% match	6%	7%	5%
85-94% match	10%	10%	7%
75-84% match	6%	7%	6%

Table 5.1 Composition of TM matches

³⁴ In using this function, 'Formatting differences penalty %' setting of the 'Penalties' tab in 'Translation Memory Options' was set to zero, so that the formatting difference was ignored and only the difference in the text was taken into account in determining the match ratio.

³⁵ This comprises of 3,597, 5,423, and 3,106 words respectively.

³⁶ This comprises of 242, 334, and 302 words respectively.

5.2.3 Machine translation using Systran

Machine translation was performed in three steps: 1) pre-processing by using Symantec's pre-processing scripts, 2) translating by Systran version 6, and 3) post-processing using Symantec's post-processing scripts. The pre-processing scripts included commands to make the ST more amenable to MT, such as protection of XML tags. The post-processing scripts included mainly commands that perform repetitive editing including the deletion of unnecessary spaces, postpositions, and personal pronouns, correction of styles and expressions, such as inappropriate endings and misuse of polite and non-polite forms, and replacement of lexical items. In using Systran, general and product-specific user dictionaries provided by Symantec were activated to ensure customised translation.

5.2.3.1 Processing of UI terms

Software manuals usually contain a number of UI terms, such as names for menus, dialog boxes, and options. The handling of such terms can be an issue in both human translation and MT. Human translation involves frequent lookup of the terminology, and MT can be problematic when the term is not properly recorded in the dictionary, or the same source word is used for multiple purposes. In real-life situations, in some cases post-editors are provided with a glossary and instructed to standardise the translation of UI terms, and in other cases they are instructed to not touch any UI terms as they will be standardised on the client's side after PE. Repairing the translation of UI terms can demand extra effort from human post-editors since it requires repeated terminology lookup. Moreover, since UI terms are one of the most important aspects of user manuals, post-editors may be distracted when they see them in the translated text. This was observed in the preliminary study and discussed in section 4.5. Despite the fact that it was clearly written in the PE guidelines distributed to participants that the UI terms need not be corrected even if they had been translated into obviously incorrect target words, post-editors changed the translation of some terms to seemingly appropriate words, even though no UI term list was provided to post-editors. This observation gives us insight into one of the possible elements that contributes to post-editors' cognitive effort, and also raises a question as to the best way to avoid such unnecessary cognitive load. What if the product terms are left in English? Would post-editors be equally/more/less distracted? In order to further examine this issue, we wished to test if

PE speed is significantly different if the UI terms are translated or left in English. In the test set for the present study, some product terms were translated, and other terms were left in English, and the post-editors were instructed not to make changes in either case. The test corpus contained 60 UI terms that are clearly marked by XML tags `<guimenuitem>`³⁷, of which 29 were translated to Japanese and 31 were left in English.

5.3 Participant Post-Editors

The experiment conducted for the present study required native Japanese-speaking professional post-editors to participate. Hiring professionals helps to increase external validity, as we can observe actual post-editors' work in the real world. This section discusses the population, sampling, and the profiles of the participants.

5.3.1 Population

The population of professional post-editors in Japan is difficult to ascertain. One reason is that it is a rather new and yet to be recognised profession, therefore no reliable survey data about post-editors have been published as far as the researcher of the present study is aware. Even when industry survey data include information about MT, post-editors are not given much attention. Therefore, we do not know how many native Japanese-speaking professional post-editors exist to begin with, though it is not difficult to estimate that the population is quite small compared to that of translators. In fact, post-editors are most often translators who take the post-editing tasks when asked.

5.3.2 Sampling method

Though random sampling is the most preferable sampling method in statistical analysis to maintain external validity in order to make generalisations, the above-mentioned situation made random sampling impossible. Therefore, the participant post-editors in this study had to be selected based on a non-random method. Among the list of non-random methods Frey et al. have listed, namely, "Convenience Sample", "Volunteer Sample", "Purposive Sample", "Quota Sample", and "Network Sample" (Frey et al. 1991: p.135), *purposive sample* seemed to be the most appropriate category for the

³⁷ This is the tag used in Symantec to indicate UI terms.

present study. Frey et al. describe purposive sampling as using “available subjects who possess the necessary characteristic” (ibid: p.135), and warn that this method can cause bias in the collected data.

The necessary characteristics of the participants in the present study were ‘a professional post-editor who is familiar with the tools used in this study.’ The subjects must also be able to participate in the experiment session during the researcher’s stay in Japan. We contacted the translation vendors in the list of Symantec’s contract vendors, asked them if they had contract post-editors who met above requirements, and recruited available post-editors.

This amounted to nine participants. Although the number of participant post-editors, nine, may be regarded as rather small for making statistical inference in terms of between-subject difference, this number may be appropriate for qualitative analysis with some depth, as Creswell & Plano Clark (Creswell & Plano Clark 2007) state that to understand the phenomenon deeply, the number of participants needs to be limited to a small number.

All participants have been chosen via Japanese translation vendors, and paid for their work based on the vendors’ regular fees. Although Symantec, the sponsor and supporter of this research, has some contract post-editors, they were excluded from this research due mainly to their non-availability at the time.

5.4 Experiment

The main data set analysed in the present study was collected from a series of experiment sessions, in which we had participant post-editors post-edit MT output, and recorded the time and the PE activity. This section describes the methods and consideration taken when conducting the experiment sessions.

5.4.1 Time and duration

The PE sessions were conducted in Japan during the period of the 16th of July 2009 to the 12th of August 2009. As discussed in section 5.2.1.2, we estimated that the planned

PE task would be feasible within approximately one day by professional translators and post-editors, thus we hired each participant for two days in order to ensure that every participant finishes the task, and would still have time left for answering the questionnaire. The PE sessions started either at 9.30 am or 10.00 am, based on the convenience of each vendor, on the first day of the allotted two-day time frame, and ended when the post-editor completed the task; all participants completed the whole task within two days. The minimum time spent on the task was 5 hours and 24 minutes, and the maximum 15 hours and 55 minutes, including lunch and other breaks.

5.4.2 Environment and equipment

Frey et al. (1991) state that the best way to achieve ecological validity, which is important in generalising the results to the real life situations, is to conduct the study in natural settings. Creswell & Plano Clark (2007) also point out that for qualitative study, a deeper understanding of the phenomena can be obtained by collecting data at the participants' actual site. In this study, all PE sessions were conducted in the office of each vendor that provided post-editors, using vendors' own computers. This was generally beneficial as it enabled the post-editors to work in their familiar environment. However, this also meant that each participant worked in a different environment depending on the conditions of the offices and the performance of the computers they used. The sites for the PE sessions at some vendors were one of their meeting rooms, where there was no one else working in the same room other than the participants, while other vendors provided a desk for each participant to work in their office where other people were working in the same room in close proximity. This could be a threat to "procedure validity and reliability" (Frey et al. 1991: p.126), which should be maintained by having consistent experimental settings for all participants. However, although this difference in environment could have an influence in between-subject design research, it may be less influential in the present study as it is mainly based on a within-subject design.

As to the computers, though there were slight differences in each machine's components, including the CPU, clock speed, RAM, display adapter, keyboard and mouse, it is difficult to believe such differences have had a significant impact on post-editors' performance as it was confirmed by the researcher that all the computers

exhibited good enough performance to use PE tools (SDL Trados Translator's Workbench and TagEditor) comfortably. In the same way as the preliminary study, the participants were asked to use these tools for post-editing, as explained in section 4.2.2.

5.4.3 Project kit

A project kit was distributed to the participant post-editors in the form of electronic files and folders. The kit included a project manual (see section 5.4.4.1), a questionnaire sheet (see section 5.5), a test kit for practising the PE task (see section 5.4.5), files to be post-edited, TM files, and setting files for Trados Translator's Workbench.

5.4.4 Instructions

In order to maintain consistency in how the experiment was conducted, written instructions and oral instructions were provided as follows.

5.4.4.1 Written instructions

An online copy of a written project manual was distributed to each vendor, which was printed and handed to each participant post-editor prior to the PE sessions. The project manual contained three sections: *Project summary*, *Task instructions*, and *Post-editing guidelines*.

The project summary contained a brief explanation of the purpose and procedures of the project as well as the policy for personal data protection. This section was provided to make it clear that the project was part of a research study, and the data obtained would only be used for research purposes.

The task instructions included the information necessary to complete the project. Firstly, a list of files and the folders provided for the project was shown with a brief explanation of each item. This gave the participants an overview of the project kit distributed to them. Secondly, steps for setting up and using various tools, namely, SDL Trados Translator's Workbench, the macro, and the screen capture software, were explained. Although Trados is the most popular TM tool among translators (Lagoudaki 2006), it was necessary to explain how to make some specific settings in Trados for this study. In addition, a macro and screen capture software were employed in this project,

which were new to the participants, thus it was necessary to explain how to use them. Finally, procedures for carrying out PE tasks and important points were emphasised. The emphasis was on the significance of certain key strokes and other operations for keeping precise time records and required file handling.

The post-editing guidelines outlined the required quality for the PE outcomes, examples of errors and problems that should be or should not be corrected, and special notes for handling UI terms. One of the biggest difficulties in compiling the PE guidelines lay in determining how much information to include. Lengthy guidelines can overwhelm post-editors with too much information and delay the process of their getting used to the job, which would affect the effective collection of time data. Overly concise guidelines, on the other hand, might leave the post-editors uninformed, which could cause them to spend too much time wondering what to do in individual cases. In this study, we aimed at following Symantec's actual PE guideline as much as possible, but needed to get the post-editors started as quickly as possible in order to collect data within a set timeframe. Therefore the PE guidelines were constructed to be as succinct as possible, based on Symantec's PE guidelines, while still emphasising the quality requirements for post-edited text; the target text had to convey correctly the meaning of the ST, and conform to Japanese grammar, but did not have to be stylistically sophisticated. The participants were requested to read the project manual before the PE session. The entire project manual in the original Japanese and its English translation is presented in Appendix A.

5.4.4.2 Oral instructions

Although the project manual was written so that the information contained in it should suffice to complete the required task, an oral instruction was also given by the researcher at each vendor's site. Oral instruction consisted mainly of a brief explanation of the purpose of the project and review of the written project manual, emphasising the important points as to procedures for obtaining valid data. For example, as explained in section 5.1.1.2, it was critical that the participants use the keyboard rather than the mouse to close the segment as that was how the keyboard macro was activated to keep the time record of every edited segment, but the importance was often not conveyed clearly enough by only the written instructions, and the detailed oral explanation of the

mechanism was often appreciated by the participants. Oral instruction took around 30 minutes each time.

5.4.5 Procedures

We established and followed the same procedure at all four vendor sites in order to reproduce the same environment as much as possible. The session consisted of the following sub-sessions in the shown order.

- 1. Setting up of the PCs to be used for PE.** This was done by the researcher before the participation of post-editors.
- 2. Brief explanation of the purpose and the objectives of the research.** This was given by the researcher.
- 3. Explanation of the procedures for PE and how to use the tools.** This was given by the researcher by reviewing the project manual and emphasising the important points that were required for the optimal data collection.
- 4. Practice of PE using the Test kit.** This was done by the participants. The researcher accompanied them, and answered the questions from the participants. Questions were mostly about the use of tools and the quality level of PE.
- 5. Signing of the informed consent form.** This is discussed in detail in section 5.4.6.3.
- 6. PE session.** The researcher normally stayed in the same room for a couple of hours to prepare for answering any questions that might arise during the session. However, she left the office confirming that the participants had become comfortable enough with the procedures without the researcher observing them.
- 7. Data retrieval.** Because of the rather large size of the data, the researcher went back to the vendors' office in the evening of the second day to retrieve data to her external hard disk.

5.4.6 Research ethics

Ethics in research, especially for those that involve human participants, is an important issue to be aware at all stages throughout the research. Frey et al. (1991: p 140) introduce three phases of research with which ethical concerns are associated, namely, those that affect “the beginning phase of research”, “the treatment of human subjects”, and “research findings and their use”. In this section, how the ethical issues for the second phase “treatment of human subjects” were addressed will be presented.

Frey et al. (ibid) suggest four categories of rules of conduct for the treatment of human participants, namely, benefits to the people being studied, protection of the participants’ privacy, provision of information and free choice, and respectful treatment of the participants. In the present study, these criteria were respected partly by the design and the course of action of the research, and partly by following the regulations of the DCU Research Ethics Committee. The research proposal was reviewed by the DCU Research Ethics Committee and approved prior to the study. The plain language statement (Appendix E), which includes a brief explanation of the study, potential risks and benefits, confidentiality of the research, and the voluntary nature of participation in the research, was distributed to the participants, and the informed consent form (Appendix F) was signed by the participants.

5.4.6.1 Potential risks and benefits

Although one of the major purposes of any research is to offer the audience of the research new knowledge based on the findings, people who cooperated to obtain such knowledge should never suffer from the study. On the contrary, they ideally should be benefited by participation in the research. Frey et al. (ibid) give an example where interviewing depressed people served to extricate them from isolation and improve their awareness. In the current study, there may be two types of benefits to the participants: direct and indirect. The direct benefit would have been monetary compensation. Since they are professionals in the field, it was appropriate to pay them the regular fee. Indirect benefits, on the other hand, may include that they can help to understand the nature of PE process, which in turn, should help to identify ways to facilitate post-editors’ work. It was hoped that they, as pioneers of the field, felt valued by participating in the study.

Frey et al. (ibid) also point out that the researcher should avoid causing participants any harm, such as stress and discomfort. In this study, one of the possible causes of harm can be the stress of having to work using partly unfamiliar tools and procedures, such as screen capture software and a custom macro. The researcher of the current study addressed this issue by explaining the required tools and procedures in as much detail as possible, and by answering any related questions in a welcoming manner. Another possible cause of harm is the stress of being observed and recorded. This is inevitable to some extent in any study that involves observation and recording. However, as a way of alleviating the stress, the purpose of the research was explained both in written form, in the project manual and the plain language statement, and orally, emphasising that the obtained data will not be used to judge the personal skills of the participants.

5.4.6.2 Participants' privacy

Frey et al. (ibid) separate two ways of protecting participants' privacy: anonymity and confidentiality. According to Frey et al., anonymity is preferable whenever possible, but anonymity can only be achieved when the researcher cannot associate the obtained data with the participants who provided them. In the current research, in which the researcher has met all the participants, full anonymity is unobtainable. That being the case, confidentiality needs to be maintained at all possible costs. Firstly, all the data from the participants were recorded under the codes, such as, Post-Editor A, Post-Editor B, and so on. Moreover, no record that associates the codes with the real identities of the participants exists in any form, including electronic or paper-based, except for the researcher's memory, which is expected to fade over time. Secondly, no personal information was gathered by the questionnaire. Although the age, education, and other demographic information may help explain some of the differences in the PE outcome, questions as to those aspects were suppressed as post-editors' personal background is not of interest in this study. More importantly, as Christians (2003) points out, real identity can be recognised by insiders even when pseudonyms are used. Considering the yet small community of post-editors in Japan, demographic information might help reveal the identity of the participants. Finally, when the data including the questionnaire sheets were retrieved by the researcher from the computers at the vendors' sites, all the

data regarding the study were erased from the computers, so that the results were kept only by the researcher.

5.4.6.3 Informed consent and free choice

Ethical researchers would agree that the participants should be provided with the full information about the purpose and the consequence of the research (Frey et al. 1991, Christians 2003). In this study, the fullest possible information was given by means of the project manual, oral instruction, question and answer session, the plain language statement, and the informed consent form. However, as is also pointed out, revealing the full information about the research can distort the results possibly because of the behavioural changes of participants caused by the knowledge about the research. In the case of the current study, since it was made clear that the time and the screen activities of PE would be recorded and used for the analysis, we cannot deny the possibilities that the participants would have behaved differently, such as trying to work more quickly than normal, or performing or not performing certain operations. The importance of recording their normal work behaviour was emphasised, but the researcher can only hope it was achieved.

At the beginning of the PE session, an informed consent form was handed to and signed by each participant. It was clearly stated in the informed consent form that their participation was voluntary and that they could withdraw from the study without any penalty, which was read by the participant and signed for the full understanding of the conditions. The researcher, however, was not involved in the hiring process of the participants by the vendors, thus having no knowledge as to how the freedom of choice in participation was achieved at that phase. Nevertheless, the researcher believes it is appropriate to assume there was no physical or psychological coercion involved in the hiring process, considering that normal business ethics do not allow such conduct.

5.4.6.4 Respectful treatment of people

Frey et al. (ibid) warn of the trap of dehumanising the subjects, and highlight the importance of treating participants as individuals to be respected. In this study, all participants were respected and valued as professionals in the field, and their questions and opinions as to the research were welcomed as worthy input. In addition, the

questionnaire sheet included a comment column, in which the participants were requested to freely express their thoughts and opinions about the research project. Frey et al. also suggest that a researcher tries to design the research so that it makes the most of the participants' valuable time and effort. In order to efficiently conduct the study, a number of preparations had preceded this study. A preliminary study was conducted several months before this study, in which the main methodology was confirmed to be valid. A project manual was distributed so that the participants could grasp the idea about the project in any convenient time for them prior to the study. The tools and the file set were installed on the computer prior to their participation as fully as possible.

5.5 Questionnaire

A questionnaire sheet was included in the project kit, which participants were asked to fill in after finishing the PE task. In this section, the aim and the design of the questionnaire used in this study is discussed. The full list of questions, both in the original Japanese and in English translation, is presented in Appendix D.

5.5.1 Aim of questionnaire

The purpose of the questionnaire in this study is two-fold. One is to gain an insight into post-editor differences that may have affected the PE process. The other one is to obtain opinions about PE and MT in general and in this project.

The questionnaire will be used to collect data for quantitative and qualitative analysis. Some of the questions have been designed to serve as independent variables in statistical analysis that may explain participant-specific differences in PE results. Other questions have been designed to obtain qualitative information to understand participants' opinions on various aspects of PE more fully.

5.5.2 Subjects and categories of questions

Different researchers have suggested different ways of categorising questions. For example, Hague (1993) classifies questions into three categories: 1) Behavioural questions, which are to acquire factual information about the respondents, such as who they are, what they do, and how often they do something, 2) Attitudinal questions, for

example, what the respondents think of or feel about something, or why they do something, and 3) Classification questions, which focus on demographic information about the respondents, such as age, gender, family composition, and so on.

In the present study, however, demographic questions were not included in the questionnaire for mainly two reasons. Firstly, although the demographic attribute of the participants could affect their PE behaviour, it is not our core interest. The primary interest of this study is to understand the ‘common causes’ for increasing PE effort among all the participant post-editors. Although the secondary interest is to find out if any participant-specific aspects affect the amount of PE effort, we are only interested in their professional conduct and not personal attributes, such as age or family compositions. Secondly, we do not have a large enough number of participants to draw a conclusion based on the demographic information.

The questions of interest in this study can thus be categorised into two major types of information, that is, facts about the participants, and the participants’ opinions. This may be closer to the classification suggested by Oppenheim (1992), which consists of 1) factual questions and 2) non-factual questions, including opinions, belief, awareness, knowledge, and attitude.

The facts about the participants we are interested in most are their experience that may affect the process and the speed of PE. Therefore, the questionnaire included a number of questions about their experience in translation and PE in related and unrelated subject domains, training in PE, and the tools and the formats used in this project. Of additional interest from a factual viewpoint is their method of PE, thus we included some questions about the steps they have taken when post-editing.

We were also interested in participants’ opinions about PE training, PE guidelines, and the difference between post-editing of MT output and editing of TM fuzzy matches. We were also interested in the usability of the tools used in this project, in order to find out if specific tools have caused additional difficulties in PE tasks. Additionally, we also included a question asking for general thoughts about this PE study project in order to learn if there have been any procedural problems in the project, and to provide opportunities for the participants to express feelings about participating in the project.

5.5.3 Designing questions

There are some common criteria suggested by researchers to be considered when designing questions, which are: the order, number, wording, and format of questions, as well as the form of questionnaire. In this section, how each of the criteria was considered in designing the questionnaire in this study will be discussed.

5.5.3.1 The order of questions

Different researchers suggest best practice about the order of questions from different viewpoints. For example, Frey et al. (1991) suggest that demographic questions should come first, while Murata et al. (2007) argue demographic questions should come last as private questions can make respondents defensive and possibly reluctant to cooperate if placed first.

Frey et al. (ibid) and Murata et al. (ibid), however, agree that the questions regarding similar topics should be grouped together. Also, Murata et al. recommend that a questionnaire should start with easy-to-answer questions and move on to the questions that require more thinking. These two are the basic approaches taken in the present study. The questionnaire in this study was designed so that it started with questions regarding participants' experience and moved onto questions about their thoughts and opinions.

5.5.3.2 Number of questions

Oppenheim (1992) points out that many surveys contain questions that have been included only for the sake of pure interest, and not for obtaining information relevant to the research. Murata et al. (ibid) and Takahashi & You (1990) suggest that the questionnaire should contain only as many questions as can be answered within approximately 30 minutes, as respondents tend to feel tired after that threshold.


The questionnaire in this study contains a total of 34 questions; 25 closed questions and 9 open questions. The answer time was expected to be around 30 minutes, but it might have taken longer depending largely on how much information respondents gave to open questions.

5.5.3.3 Wording of questions

Researchers agree on some of the important points regarding wording of questions; questions need to be unambiguous and easy to understand, cover only one issue at a time, avoid directing respondents to answer in certain ways, and be neutral by avoiding emotional words and expressions (Takahashi & You 1990, Frey et al. 1991, Oppenheim 1992). These suggestions were taken into consideration when designing questions in the present study. For instance, questions about the participants' experience were organised so that each of them enquires about only one aspect of experience at a time regarding translation and/or PE in the related and/or unrelated subjects and using the tool employed in this study.

5.5.3.4 Questionnaire form

The questionnaire for this study was distributed in the form of an MS Word (.doc) file utilising its Form feature, instead of printed paper. The potential disadvantages of using electronic forms as opposed to paper in general include the possible loss of data and the difficulty for respondents in using the word processing software. The loss of data may occur by having the file accidentally deleted by the respondents or researcher, or being lost at some stage of transferring the data. To avoid such accidents, the data were duplicated on more than two different hard disks. The second disadvantage may not be an issue in this study as all respondents are currently working professional translators, who constantly use computers for daily work and are proficient in using standard business software. In such a case, the electronic form could offer an advantage as it gives respondents ease of ticking and un-ticking the options for closed questions as well as writing, deleting, and editing the answers for open questions to make them as structured and meaningful as possible.

Oppenheim (1992) points out that as for open-ended questions, the space allocated to the question partly determines the amount and the fullness of the information one can expect to receive. The Form feature of MS Word gives a specific space in the form of a grey box for open questions, which seems to accommodate only several characters (ex. []), but in reality expands as the text is typed in. In order to give the respondents the idea that they are expected to write as much as they want, the following comment was inserted at the beginning of the questionnaire:

For free-answer questions, please click the grey box within the brackets and start typing. The space in the brackets expands as you type. Please answer in as much detail as possible.

For the full questionnaire both in original Japanese and English translation, see Appendix D.

5.6 Concluding Remarks

This chapter discussed the practical methods employed in the present study. The chapter began by explaining the techniques for measuring each variable incorporated in the research questions and how we addressed measurement reliability. It then outlined the steps taken in each phase of the study, including compiling the test corpus, recruiting participants, conducting the experiment, and creating the questionnaire, and the rationale and logic of employing certain methods were explained, highlighting how the methodological framework was transformed into practical methods so that the research questions can be answered while maintaining the expected levels of rigor in both quantitative and qualitative aspects of the study.

PART III
RESEARCH FINDINGS AND DISCUSSIONS

Chapter 6 TAXONOMY DEVELOPMENT FOR PE OPERATION TYPES

As discussed in Chapter 3 (section 3.2), the first part of the analysis is carried out using the exploratory design, which consists of two phases: 1) qualitative analyses of PE operation types for taxonomy development, and 2) quantification of the findings from the first phase. In this chapter, however, only the qualitative analysis phase (Figure 3.1, label [2]) will be discussed, leaving the quantification phase to Chapter 7. This chapter first explains the method of qualitative analysis employed in this study in section 6.1, and presents the findings from the analysis in section 6.2.

6.1 Method of Qualitative Analysis

Hughes & Hayhoe (2007) explain the three phases involved in qualitative data analysis: coding, categorising, and modelling. In the present study, however, since the qualitative analysis will be done in order to develop a PE classification system, only the coding and categorising phases will be concerned. The following is the summary of their explanations of each phase.

Coding consists of two steps: breaking down the data into the desired unit of analysis, and recounting the data at the unit level. In this phase, the researcher works closely with the data. A researcher can use *predefined codes* when an appropriate code set is already available, but it is normal to use *open codes* when a new pattern is to be discovered in the study, in which case the characteristics of data units need to be directly gleaned from the collected data. Open codes are fluid, and often need to be modified along the way while establishing the codes, which can be challenging, but necessary in order to draw insight from the data without preconceptions. In the present study, open codes were used for the reasons discussed in section 3.4.4.

Categorising involves finding out the patterns or groups in the coded data and classifying each piece of data into categories. By working from data to codes to categories, this phase makes it possible to abstract the findings so that the analysis can be applicable to data other than those observed in the current research.

6.2 Taxonomy Development for PE Operation Types

6.2.1 Coding

As mentioned in Chapters 3 and 4, we employed some automatic measures to start the exploratory phase of analysing PE operation types to have a consistent basis for analysis.

It is difficult to identify in a standardised manner what has been done to transform one sentence to another if a human looks at the two sentences and tries to spot the changes, since often one or more word change(s) and structural change(s) are made in a sentence, which may obscure which part of the edited sentence corresponds to which part in the MT output. In addition, the Japanese writing system does not insert spaces to distinguish a word from another, as mentioned in section 5.1.2.1, which makes it even more difficult to isolate each edit instance. It is more manageable if the sentence is chunked into smaller units and the parts that have been changed are highlighted in a systematic manner. When Abekawa & Kageura (2008b) tried to identify English to Japanese human translation revision patterns, they employed GIZA++, a word alignment program, to match morphs in the draft translation with those in the revised translation so that they can make comparisons between corresponding pairs of morphs. We employed a method that is theoretically similar to theirs, but using different tools. In the present study, coding was performed in three steps: 1) chunking the sentences using a software program, 2) highlighting the edited chunks using another software program, and 3) assigning each edited chunk a PE operation type manually.

6.2.1.1 Coding step 1: Chunking and part of speech tagging

First of all, all MT output and PE output was chunked or tokenised into morphs by means of MeCab, a Japanese morphoanalyser and tokeniser mentioned in section 5.1.2.1. This process was necessary in order to split Japanese text into edit units, as stated earlier. MeCab not only tokenises the text, but also assigns a part of speech to each morph, which gives insight into which parts of speech are related to what PE operation. As the IPA dictionary³⁸, one of the standard Japanese dictionaries used by MeCab, and the one we employed, classifies each morph into highly granular part of

³⁸ IPA dictionary was constructed based on the part of speech structures established by Information-Technology Promotion Agency.

speech categories, we aggregated some of the categories for simplification, borrowing the convention from a Japanese grammar dictionary (Hayashi et al. 2004), and translated into equivalent English terms. The full list and examples of the part of speech categories we employed in the present study are given in Appendix B.

6.2.1.2 Coding step 2: Identifying the foci of interest and technical effort type

We then needed to highlight the parts that have been modified during PE. For this purpose, TER, one of the AEMs, was used to find out which morphs were post-edited. This step not only identified what morphs were modified during PE, but also assigned the type of ‘technical effort’ (Krings’s terminology discussed in section 2.3) of each PE operation in a systematic manner. TER does this according to its own algorithm, and categorises the changes into four edit types: Insertion, Deletion, Substitution, and Shift.

As a result of steps 1 and 2, the researcher was provided with the list of post-edited morphs with their parts of speech and technical effort types, of which an example is shown in Table 6.1.

MT output:

適切なアプリケーションのコンテンツを開くためにショートカットをダブルクリックします。

PE output:

ショートカットをダブルクリックすると、適切なアプリケーション内でコンテンツが開きます。

MT output	(Part of Speech)	PE output	(Part of Speech)	PE type
ショートカット	(Noun)	ショートカット	(Noun)	Shift: -9
を	(Case particle)	を	(Case particle)	Shift: -9
ダブル	(Noun)	ダブル	(Noun)	Shift: -9
クリック	(Noun)	クリック	(Noun)	Shift: -9
		する	(Sa-Verb)	Insertion
		と	(Conjunction particle)	Insertion
		、	(Punctuation)	Insertion
適切	(Noun)	適切	(Noun)	
な	(Auxiliary verb)	な	(Auxiliary verb)	
アプリケーション	(Noun)	アプリケーション	(Noun)	
の	(Case particle)			Deletion
を	(Case particle)	内	(Suffixal noun)	Substitution
開く	(Verb)	で	(Case particle)	Substitution
コンテンツ	(Noun)	コンテンツ	(Noun)	Shift: 2
ため	(Dependent noun)			Deletion
に	(Case particle)	が	(Case particle)	Substitution
し	(Sa-Verb)	開き	(Verb)	Substitution
ます	(Auxiliary verb)	ます	(Auxiliary verb)	
。	(Punctuation)	。	(Punctuation)	

Table 6.1 An example of aligned information extracted from TER and MeCab

The first and the second column show the breakdown of MT and PE output with the part of speech of each morph respectively. The third column shows the type of edits performed to convert the MT output to PE output. When a deletion has been performed, there is an entry only in the first column, and when an insertion is performed, there is an entry only in the second column. There is an entry in each of these columns when a substitution has been performed, while entries in both columns are the same when the morph itself has not been changed but only its position has been changed, that is, a ‘shift’ has occurred. The number in the third column after ‘Shift’ shows the distance of the move. For example, the first four entries in the first column, ‘ショートカット’, ‘を’, ‘ダブル’, and ‘クリック’, had been positioned after ‘に’ in the original MT output shown above the table (dotted underline), and were moved backwards a distance of nine morphs, thus shown as ‘Shift: -9’ in the third column.

These aligned data gave us information on how many insertions, deletions, substitutions, and shifts have been performed on what parts of speech, and also on which specific morphs, which is used as the basis for finding out the pattern of PE. The advantages of this method include that it eliminates the overlap of categories, as a morph and a part of speech have a one-to-one correspondence, avoids inconsistency since the classification is carried out by software programs, and can be automated by combining the functions of two programs. The aligned list served as the starting point for qualitative analysis and guided the researcher to find out what happened to the text during PE, based on which the third step was performed.

6.2.1.3 Coding step 3: Describing the effect of PE operation

This step was performed manually by the researcher of the present study. This step was based on the information obtained from the steps 1 and 2: part of speech and technical effort type, and by establishing open codes in order to gain insight into PE operation types directly from the data.

We first looked at the *bunsetsu*³⁹ level changes, which we define as any change within one *bunsetsu*, or between adjacent *bunsetsu*. We will call this type of change a ‘*bunsetsu level*’ edit in this study. By looking at the *bunsetsu* level changes, we noticed that there are different effects that *bunsetsu* level PE brings about, which eventually converged into three types: supplementation, omission, and alteration. A *supplementation* occurs when 1) a concept that does not exist in the MT output is added in the PE output, or 2) a pronoun in MT output is replaced with the actual noun that is referred to by the pronoun. One of the examples of this is the conversion of an anaphora into the actual content word. On the other hand, an *omission* occurs when a concept that exists in MT output is omitted in PE output. One of the examples of this is again an anaphora being omitted when it is obvious without mentioning it, which is a pragmatic feature in the Japanese language. An *alteration* occurs when one or more morph(s) that form one concept is/are changed to other morph(s) to alter the meaning. Alteration can be divided firstly into two groups: semantic changes and non-semantic changes.

³⁹ There are a number of definitions for the Japanese linguistic unit ‘*bunsetsu*’. In the present study, however, we employ Yoshimura’s definition: *Bunsetsu* is the smallest unit resulting from dividing a sentence in a semantically reasonable and natural way. A *bunsetsu* consists of one content word/morph and zero or more function word(s)/morph(s) (Yoshimura 2000 p.2).

Semantic changes can be further divided into three groups depending on the type of change in relation to terminology: UI terms, Technical terms, and General terms. Non-semantic *bunsetsu* level alterations are all classified into the category ‘*Bunsetsu* level stylistic change’.

The combination of the part of speech and the edit type sometimes indicates which type of PE operation has been performed. For instance, a supplementation often happens when a content morph is inserted.

(For all the following examples, ‘ST’ means the English source text, ‘MT’ machine translation output, and ‘PE’ post-edited text. The relevant part is highlighted by an underline; only the relevant part is considered in each example. The gloss for the applicable part is shown in the square brackets at the end of MT and PE.)

ST: You can give other users access to your folders ...

MT: ... フォルダへの他のユーザーアクセスを与えることができます。 [access]

PE: ...ユーザーにフォルダへのアクセス権を与えることができます。 [access rights]

A supplementation can also happen as a result of substituting a pronoun with a corresponding general noun.

ST: You must have the Folder Full Control role in the folder to give other users access to it.

MT: それへの他のユーザーアクセスを与えるフォルダのフォルダのフルコントロールのロールを持たなければなりません。 [it]

PE: フォルダへの他のユーザーアクセスを与えるにはそのフォルダのフルコントロールのロールを持たなければなりません。 [folder]

On the other hand, when a pronoun and a following case particle are removed, it is called an omission.

ST: ... and adds them to the review set.

MT: レビューセットにそれらを追加します [them]

PE: レビューセットに追加します [φ]

A substitution from one content morph to another content morph of the same part of speech is likely to be related to a *bunsetsu* level alteration. The following example contains one instance of noun alteration and one verb alteration:

ST: ... filter option shows the number of matching items

MT: ... フィルタオプションは一致したアイテムの番号を 現します [put into sight, the sequence number]

PE: ... フィルタオプションは一致したアイテムの数を 表します [indicate, the quantity]

There are also changes that span a broader scope than the *bunsetsu* level. We will call this type of change a ‘*structure level*’ edit in this study. A typical example of this is when a group of adjacent *bunsetsu* have been shifted. In some cases, especially one or more content morph(s) with an adjacent case particle or conjunctive particle have been moved, this can involve a change in the relationship between two or more *bunsetsu*, resulting in a dependency edit. For example, if one or more content morph(s) are moved with a possessive case particle, it is likely to result in a change in the relationship between a modifier and the modified morphs.

ST: ... show data ingestion progress, and the status of the automatic categorization.

MT: ... 自動類別のデータ取り込みの進行状況とステータスを現します。 [...show data ingestion progress of the automatic categorization and the status]

PE: ... データ取り込みの進行状況と自動類別のステータスを現します。 [...show data ingestion progress and the status of the automatic categorization]

A shift is sometimes performed not in order to change the meaning, but perhaps in an effort to enhance naturalness. For example, when the positions of two verb modifiers are swapped, it is likely to result in just a structural change and does not involve a change in the meaning of the sentence.

ST: You can add the employees that you want to monitor to the appropriate departments.

MT: 適切な部門に 監視する従業員を追加できます。 [You can add to the appropriate departments the employees that you want to monitor.]

PE: 監視する従業員を 適切な部門に追加できます。 [You can add the employees that you want to monitor to the appropriate departments.]

The above mentioned are a few of the examples where the combination of part of speech and edit type can be linked to a PE operation. However, since this does not always happen, the information as to part of speech and edit action alone is not sufficient for identifying the PE operation type; it was necessary for the researcher of the present study to go through all the morphs and groups of morphs manually to qualitatively examine what type of PE operation has been performed on each morph.

As Hughes and Hayhoe (2007) cautioned, this was not a straightforward process, which involved a number of experimental coding, re-coding, and dividing and joining of temporarily defined codes, which eventually boiled down to ten translation-related codes and two non-translation-related codes, which are described with some examples below.

Code 1: Supplementation

As mentioned previously, supplementation is the act of adding one or more morphs to supplement the concept. In the following example, the English words ‘ever’ and ‘found’ are not explicitly translated by MT, and supplemented during PE:

ST: *If these items are ever found by another search,*

MT: *これらのアイテムが別の検索によってあれば,* [If these items exist by another search,]

PE: *これらのアイテムが過去に別の検索によって見つかったことがあれば,* [If these items have been found by another search in the past,]

Another example shows the case where the post-editor added some morphs to complement the insufficiency of the direct Japanese translation of the English word ‘following’ and ‘does’:

ST: *When you perform a new search and accept the results, XXX does the following for each item that you accept:*

MT: 新しい検索を実行し、結果を受け入れるとき、XXX は受け入れる各アイテムのための次をします。 [*does the following*]

PE: 新しい検索を実行し、結果を受け入れるとき、XXX は受け入れる各アイテムに対して以下の作業を実行します。 [*performs the following task*]

A third example is to avoid using an anaphora and repeat the referred content morph(s). For example:

ST: *You can give other users access to your folders by assigning roles to them.*

MT: それらにロールを割り当てることによってフォルダへの他のユーザーアクセスを与えることができます。 [*them (translated into inanimate pronoun)*]

PE: 他のユーザーにロールを割り当てることによって、他のユーザーに自分のフォルダへのアクセス権を与えることができます。 [*other users*]

Code 2: Omission

An omission is the act of removing information that exists in the MT output. This is usually not related to accuracy, but performed for the sake of stylistic naturalness in Japanese. A typical example is the omission of a pronoun. The following is an example where a subject pronoun is omitted.⁴⁰

ST: *It also checks for existing marks.*

MT: それはまたマークを存在することをがあるかどうかを調べます。 [*It*]

PE: またマークが存在するかも調べます。 [*φ*]

An object pronoun is also often omitted as the following example shows.

ST: *Let us know what you like and dislike about the documentation.*

MT: 私達にマニュアルについて好み、嫌っているものを知らせてください。 [*us*]

PE: マニュアルについて、気に入った点と気に入らない点をお知らせください。 [*φ*]

⁴⁰ Japanese sentence does not require a subject (Hayashi et al. 2004).

Another example is to omit the adverb ‘then’, when it does not play a meaningful role or causes confusion in the MT output. This omission is common practice in Japanese technical translation (Kosaka 2002).

ST: You can then use this name as a shorthand way of referring to the list of people.

MT: 人々のリストを参照する速記の方法としてそれからこの名前を使うことができます。

PE: ユーザーリストを参照する簡単な方法として、この名前を使うことができます。[φ]

Code 3: UI Term Alteration

This includes menu names, option names, and messages that appear on the product interface and are clearly marked by <guimenuitem> XML tags as discussed in section 5.2.3.1. These terms need to match the product-specific terminology. UI terms need to be treated in a specific way depending on the company policy. In the case of the present study, the post-editors were told not to make any modifications to UI terms even if they had been left in English or translated into seemingly inappropriate Japanese. Nevertheless, a number of changes have been made during PE by a number of post-editors; Japanese terms were often changed either into transliteration or back into English. This had also been observed in the preliminary study and we were interested in finding out if the sentences that include UI terms significantly slow down the PE process. An example of UI term alteration is:

ST: Click <guimenuitem> New </guimenuitem> , and then click <guimenuitem> Mark </guimenuitem> .

MT: <guimenuitem> 新規 </guimenuitem> をクリックし、 <guimenuitem> EP </guimenuitem> をクリックします。[mark (in the general sense)]

PE: <guimenuitem> 新規 </guimenuitem> をクリックし、 <guimenuitem> マーク </guimenuitem> をクリックします。[mark (as a symbol for indication)]

Code 4: Technical Term Alteration

This includes the terms that are product specific but not marked as UI terms and the technical terms that are standardised in Symantec. These terms are listed in Symantec’s user dictionary for Systran, thus in theory are expected to be translated correctly by Systran. For this reason, the post-editors were not provided with any Symantec glossary.

This is in line with the TAUS's definition of light post-editing, which assumes that the terminology is already covered in the MT dictionaries, thus "almost no in-depth term checking is required." (TAUS 2010: p.8). Since no glossaries were provided, it was not possible for the post-editors to precisely distinguish the company-specific technical terms from general terms, though in many cases the difference may be obvious to the eyes of experienced translators and post-editors. In either case, we were interested in finding out if any significant difference in PE speed could be observed between editing technical terms and general terms. Some of the examples are:

ST: Users with the application-wide Delete Department permission can delete any department in the system.

*MT: アプリケーション全体の削除の部門権限のユーザーはシステムの部門を削除できます。
[department of deletion (incorrectly translated by Systran)]*

PE: アプリケーション全体にまたがる部門の削除権限を持つユーザーは、システム内の任意の部門を削除できます。 [deleting of departments (correctly edited by the post-editor despite the lack of access to the glossary)]

Sometimes a term whose translation is registered in a user dictionary may not be perceived as correct by the post-editors and consequently the translation is modified unnecessarily.

ST: If you carry out a production run and then ...

MT: 提出物生成実行を実行し、 [deliverable production run (correctly translated by Systran)]

PE: 生成実行を実行した場合に、 [production run (incorrectly modified by the post-editor)]

Code 5: General Term Alteration

This includes any general lexical items that are not classified into UI or Technical terms. Some of the examples are shown below:

ST: For example, you may want a mark that is called Spam to stay with items in the master collection.

MT: たとえば、マスターコレクションのアイテムととどまるスパムと呼び出されるマークがほしい場合もあります。 [summoned]

PE: たとえば、スパムと呼ばれるマークをマスターコレクションのアイテムと関連付けするとします。[named]

ST: Then you can specify the target name in your search criteria as a shorthand way of listing all the associated addresses.

MT: それからすべての関連付けされたアドレスをリストする速記の方法として検索基準でターゲット名を指定できます。[a stenographical way]

PE: それからすべての関連付けされたアドレスをリストする簡単な方法として検索基準でターゲット名を指定できます。[an easy way]

ST: Later, you can refine the list of associated roles when you customize the marks that are associated with an individual case.

MT: 個々のケースによって関連付けされるマークをカスタマイズするとき後で、関連付けされたロールのリストを精製できます。[purify]

PE: 後で個々のケースに関連付けられているマークをカスタマイズするとき、関連付けされたロールのリストをさらに細かく設定できます。[more finely define]

ST: Some items in the master collection are also in cases A and B.

MT: マスターコレクションのあるアイテムはケース A と B. にまたあります。[a certain]

PE: マスターコレクションの一部のアイテムはケース A と B にもあります。[some of the]

ST: Users who occupy these roles can apply this mark to the items that they review.

MT: これらのロールを占めるユーザーは確認するアイテムにこのマークを加えることができます。[cover entirely]

PE: これらのロールを持つユーザーは、確認するアイテムにこのマークを適用することができます。[hold]

Code 6: *Bunsetsu* Level Stylistic Change

This is similar to the code General Term Alteration, in that this is limited to general terms, but changes do not affect the meaning of the term. It was often observed that when Systran produced expressions in general or casual languages, they were often corrected by post-editors into a more formal style language, which is more in keeping with a standard technical writing style. For example:

ST: *XXX keeps a master collection of all items that have ever been accepted into any case.*

MT: XXX はあらゆるケースに受け入れられたあることがすべてのアイテムのマスターコレクションを保ちます。[keep]

PE: XXX はあらゆるケースに受け入れられたすべてのアイテムのマスターコレクションを保持します。[keep/hold]

ST: *If the mark does not have this property, the mark applies within the case but is not available to other cases.*

MT: マークにこのプロパティがなければ、マークはケースの内で適用しますが、他のケースに利用可能ではありません。[if this property does not exist for the mark,]

PE: マークにこのプロパティがない場合、マークはケース内で適用されますが、他のケースには使用できません。[in the case that this property does not exist for the mark,]

In other cases, MT translation is acceptable, while there is an option for another expression, which is more commonly used in IT-related documents. The following example shows the case where a Japanese verb is changed to a transliteration. Itagaki et al. (1999) discuss that transliteration is a widely used technique in the computer industry as it has benefits that are especially advantageous in the industry, including the fact that it makes it clear that the concept is computer-specific, it avoids mistranslation, and it makes it easy to guess the corresponding English term, which is convenient as a great deal of computer-related reference materials are available in English.

ST: *It also checks for existing marks.*

MT: それはまたマークを存在することをがあるかどうかを調べます。[check / examine]

PE: マークがすでに存在するかどうかもチェックします。[check]

As Kosaka (2002) points out, the Japanese translation of the English expression ‘want to’ often does not sound appropriate in technical text, and is preferably changed to a more neutral expression:

ST: In the right box, select the users to whom you want to assign the role.

MT: 右のフィールドで、ロールを割り当てたいユーザーを選択します。 [users to whom you want to assign the role]

PE: 右のフィールドで、ロールを割り当てるユーザーを選択します。 [users to assign the role]

Another example is dropping the possessive case particle that converts the preceding noun into an adjective. The possessive case particles tend to appear too frequently in technical translations, and it is often recommended to omit them when possible (Itagaki et al. 1999, Kosaka 2002). The following example converts an adjective-noun combination into a compound noun without affecting the meaning:

ST: Sampling mode

MT: サンプルングのモード [mode of sampling]

PE: サンプルングモード [sampling mode]

Code 7: Dependency Edit

In the present study, a dependency edit is defined as the change in the relationships between *bunsetsu* as discussed in section 4.4.4. One of the common causes that necessitate dependency edits is the incorrect parsing of prepositional phrases by MT.⁴¹ An example of this is the misinterpretation of ‘with’.

ST: Defining email targets with Address Manager

MT: アドレスのマネージャが付いている電子メールのターゲットの定義 [(Defining email targets to which Address Manager) is attached]

PE: アドレスマネージャを使った電子メールのターゲットの定義 [(Defining email target) by using (Address Manager)]

⁴¹ Parsing prepositional phrases has been regarded as one of the major challenge in MT (Mamidi 2004, Gustavii 2005, Wu et al. 2006).

The ‘Address Manager’ had modified ‘email targets’ in MT, and was corrected to modify ‘defining’ during the PE. Another example is the misinterpretation of ‘to’.

ST: In the Review pane, select the items to review.

MT: レビューペインで、確認するためにアイテムを選択します。[select the items in order to perform review]

PE: レビューペインで、確認するアイテムを選択します。[select the items which you are going to review]

The ‘to review’ had modified ‘select’ in MT, and was corrected to modify ‘the items’ during the PE. There are various other causes for dependency edits. For example, an object is sometimes mistakenly interpreted as a modifier by MT:

ST: You can give other users access to your folders by assigning roles to them.

MT: それらにロールを割り当てることによってフォルダへの他のユーザーアクセスを与えることができます。[give other types of user access rights to folders]

PE: 他のユーザーにロールを割り当てることによって、他のユーザーに自分のフォルダへのアクセス権を与えることができます。[give other users access rights to your folders]

The ‘users’ had been treated as a modifier for ‘access’ to mean ‘user access’ in MT, which was changed to a direct object of the verb ‘give’ in the PE result. Another example is an inappropriate addition of a possessive suffix, which resulted in a dependency error:

ST: You can export individual items multiple times, but you can produce items once only.

MT: 個々のアイテムの複数の時間をエクスポートできますアイテムしか生成な一度できます。[export the multiple durations of individual items]

PE: 個々のアイテムを複数回エクスポートできますが、生成は 1 回しかできません。[export individual items more than once]

In MT, ‘items’ mistakenly modifies ‘multiple times’ and ‘multiple times’ is also mistakenly treated as an object of the verb ‘export’. In the PE result, both problems are corrected so that ‘item’ and ‘multiple times’ are changed to an object and a modifier of

the verb ‘export’, respectively. There are also cases where the ‘that’ clause is not translated correctly. For example:

ST: The automatic categorization summary shows the actions that the rules applied.

MT: 自動類別の概略はその処理に適用されるルールを現します。 [rules that are applied to the actions]

PE: 自動類別の概略は、ルールの適用対象の処理を表示します。 [the actions that the rules applied]

The last example is a word order problem that results in a dependency error.

ST: Maximum age of unresolved items (hours)

MT: 未解決アイテム (時間) の最大の経過時間 [(‘hours’) modifies ‘items’)]

PE: 未解決アイテムの最大の経過時間 (時間) [(‘hours’) modifies ‘age’ to indicate that hours is used as a unit of measurement for this option)]

Code 8: Rewrite

In the present study, a PE operation that involves modification of the sense and is beyond *bunsetsu* level or dependency change is classified in this group. Sometimes more than one dependency correction in one sentence can lead to a radical rewriting. A rewrite can occur on a phrase, clause, or sentence level. For instance:

ST: The roles that you assign to employees determine what they can access and the tasks that they can perform in XXX.

MT: 従業員に割り当てるロールはいいXXX で実行してもいいタスクものに判断し、アクセスしても。 [(the entire sentence is completely unintelligible and is impossible to translate back into English)]

PE: 従業員に割り当てるロールによって、アクセスできる対象とXXX で実行可能なタスクが決まります。

In other cases, only parts of the sentence are subject to a rewrite. For instance,

ST: If you use the same output folder and production run name for multiple runs, the report summary is overwritten each time.

MT: 同じ出力のフォルダを使い、複数の提出物生成実行が名前実行する場合は、レポート概略はいつも上書きされます。[gloss not provided due to the unintelligibility]*

PE: 複数の実行で同じ出力フォルダ名と生成実行名を使用する場合は、レポート概略はいつも上書きされます。

Code 9: Structure Level Stylistic Change

When the edit spans the range beyond the *bunsetsu* level but does not involve changes in sense, it was classified as a structure level stylistic change (When *bunsetsu* has been moved to a new location and that consequently changes the relationship between *bunsetsu*, it is classified as a dependency edit, but if it does not change the relationship between *bunsetsu*, it is classified as a structure level stylistic change). Though we call this category ‘Structure-level *stylistic* change’ for the sake of convenience, it is not limited to style, and includes the edits that make changes in structure, nuance, standpoint, etc. of the sentence or a phrase within a sentence. There are various cases that fall into this category. The following are some of the examples.

Some of the word order changes do not result in a change in the meaning, but seem to be done simply in order to enhance the naturalness and clarity of the sentence.

ST: When configuring an employee profile, you can choose whether XXX should automatically synchronize these properties with the corresponding directory account information.

MT: 従業員プロフィールを構成するとき、XXX が対応するディレクトリのアカウント情報と自動的にこれらのプロパティを同期するべきであるかどうか選択できます。[... synchronize these properties automatically with the corresponding...]

PE: 従業員プロフィールを構成するとき、XXX が対応するディレクトリのアカウント情報とこれらのプロパティを自動的に同期するべきであるかどうか選択できます。[automatically synchronize these properties with the corresponding...]

When it comes to the title of a section, a nominalisation using a verb form ending is sometimes further changed to a noun form ending for a more succinct expression in keeping with a technical writing style.

ST: Filtering the items in the Review pane

MT: レビューペインのアイテムをフィルタ処理すること [filtering (noun-style ending of a verb 'filter')]

PE: レビューペインのアイテムのフィルタ処理 [filtering (noun for 'filtering')]

When an instruction contains an action (method) and its purpose, the direct translation from English to Japanese produced by MT sometimes places an emphasis on the wrong part, for example, putting emphasis more on the method when purpose is more important. In such cases, PE may be performed to shift the focus of the sentence. For instance:

ST: *If you do not want the original owner to retain these permissions, use the Role Assignment facility to deassign them.*

MT: 元の所有者にこれらの権限を保持してほしいしなければそれらを割り当てから外すのにロールの割り当て機能を使います。 [use the Role Assignment facility in order to deassign them]

PE: 元の所有者にこれらの権限を保持してほしいなければ、ロールの割り当て機能を使用してそれらを割り当て解除します。 [deassign them by using the Role Assignment facility]

In Japanese, an inanimate subject is not as frequently used as in English, and thus it is common practice to change it to an unspecified human subject (Itagaki et al. 1999, Kosaka 2002). For example:

ST: *When you enable a case for analytics, this pane lets you monitor the process of ingesting the data into a case.*

MT: 分析のためのケースを有効にすると、このペインはケースにデータを摂取するプロセスを監視することを可能にします。 [this pane lets you monitor the process of ingesting the data into a case]

PE: 分析用にケースを有効にすると、このペインによって、ケースへのデータの取り込みプロセスを監視できます。 [you can monitor the process of ingesting the data into a case by using this pane]

Although this type of change does not affect the meaning of the sentence, it greatly affects the naturalness of the Japanese sentence. In fact, for this specific case, seven out

of nine post-editors shifted the subject. There are more subtle edits that make slight changes in nuance. For instance:

ST: *However, you can create and edit rules after the case has been enabled for analytics.*

MT: *分析のためのケースが有効になった後、ルールを作成し、編集できます。 [you can create and edit rules after ...]*

PE: *分析のためのケースが有効になった後でルールを作成し、編集することもできます。 [you are also able to create and edit rules even after ...]*

ST: *When you perform a new search and accept the results, XXX does the following for each item that you accept:*

MT: *新しい検索を実行し、結果を受け入れるとき、XXX は受け入れる各アイテムのための次をします。 [When ...]*

PE: *新しい検索を実行し、結果を受け入れると、XXX は受け入れるアイテムごとに以下を実行します。 [If...]*

Code 10: Punctuation Edit

An addition, deletion, or repositioning of a punctuation mark can change the meaning of the sentence, or raises the understandability or naturalness of the sentence highlighting the main idea the sentence conveys. When the ST does not contain commas, RBMT may fail to insert it automatically, thus requiring human editors to insert one. For example:

ST: *Any searches that use the following schedules run automatically at the interval that you specify in the schedule.*

MT: *次のスケジュールを使うどの検索でもスケジュールで指定する間隔で自動的に実行します。*

PE: *次のスケジュールを使うどの検索でも、スケジュールで指定する間隔で自動的に実行します。*

Code 11: Tag

This is one of the non-translation issues. When the post-editor inserted/deleted/moved XML tags, that operation was classified as a Tag edit.

Code 12: Comment

This is another non-translation issue. Post-editors were asked to insert within the target sentence a short comment when they had any questions or problems which could not be resolved by referring to the post-editing guidelines.

6.2.2 Categorising

As mentioned at the beginning of this chapter, the second phase of the qualitative analysis, according to Hughes & Hayhoe (2007), is *categorising*. In the current study, the categories of the codes were developed during the process of coding, with two phases happening in parallel. Table 6.2 summarises the categories to which each code of PE operation mentioned in the previous subsection is classified.

Translation issues:

	<i>Bunsetsu</i> level	Structure level
Sense	- Supplementation - Omission - UI Term Alteration - Technical Term Alteration - General Term Alteration	- Dependency Edit - Rewrite
Expression	- <i>Bunsetsu</i> Level Stylistic Change	- Structure Level Stylistic Change - Punctuation Edit

Non-translation issues:

- Tag
- Comment

Table 6.2 Taxonomy of PE operation types

There are seven codes that are categorised as Sense PE, five of which are *bunsetsu* level and two structure level. A Sense PE changes the meaning of the text, while an Expression PE does not. There are two Non-translation PE codes: Tag and Comment. The differentiation between Sense and Expression somewhat corresponds to the popular dichotomy for MT quality evaluation: *adequacy* and *fluency*⁴², respectively. However,

⁴² ‘Adequacy’ accounts for how well the information contained in the source text and the model translation, if applicable, has been accurately conveyed in the MT output, while ‘fluency’ only concerns the well-formedness and understandability of the target text (Turian et al. 2003, LDC 2005, Boitet et al. 2006, Snover et al. 2006, Callison-Burch et al. 2007). Similar concepts have been expressed by different terms, such as, ‘informativeness’ and ‘intelligibility’ (ALPAC 1966), ‘accuracy’ and ‘intelligibility’ (Arnold et al. 1993), and ‘fidelity’ and ‘intelligibility’ (White 2003).

determining adequacy and fluency requires judgement of PE results, which is not the focus of the present study; our focus is on what ‘changes’ are made during PE. To this end, we seek to describe the differences between MT output and PE results in a manner that is as objective as possible.

From our qualitative analysis, we suggest that some of the PE operations could be automated or avoided while others may not be. For *bunsetsu* level edits, PE of UI and Technical terms may be reduced by tagging the terms for more successful translation and/or further refining user dictionaries for RBMT systems, while it may be more difficult to automate Supplementation and Omission, especially for RBMT systems. As for structural edits, considering many of the causes for Dependency edits are mistranslations of prepositions, an improvement of MT quality for prepositions can be expected to reduce the amount of PE effort in this category. On the other hand, punctuation edits sometimes involve understanding of subtle differences in nuance caused by the placement of punctuation in each context, which may make it difficult for automation.

6.3 Concluding Remarks

This chapter first outlined the method of qualitative analysis employed in the present study. Then it moved onto the taxonomy development phase in which PE operation type categories were established. The coding phase involved three steps, namely, chunking and part of speech tagging of the text, identifying the foci of interest and technical effort types, and describing the effect of the PE operation. The categories resulting from this process will be used in the statistical analysis in Chapter 7 in order to find out what PE operation types have a high impact on PE speed.

Chapter 7 QUANTITATIVE ANALYSIS

This chapter focuses on the quantitative analysis to answer the core research questions (Figure 3.1, labels [3], [4], and [5]). Section 7.1 describes the analysed data set, following which section 7.2 summarises the quantitative data for the dependent variable and the primary independent variable. Then section 7.3 provides the answer to RQ1 [How does the amount of editing correlate with the amount of effort in post-editing of English to Japanese MT output?] by plotting the correlation between the amount of editing and the PE speed. Section 7.4 details and quantifies the independent variables in the ST characteristics category chosen for answering RQ2 [What characteristics of the English source text have significant influence on the amount of PE effort irrespective of individual traits of post-editors?]. Section 7.5 quantifies the results from Chapter 6, in preparation for answering RQ3 [What types of PE operation have significant influence on the amount of PE effort irrespective of individual traits of post-editors?]. Section 7.6 presents the results of multiple regression analyses that provide the combined answers to RQ1, RQ2, and RQ3, while section 7.7 compares the editing tasks between MT output and TM fuzzy matches to answer RQ4 [What are the differences between editing of TM match segments and post-editing of MT output?]. Finally, section 7.8 summarises the findings from the quantitative analysis.

7.1 Profile of Analysed Data

We already mentioned how we compiled the test corpus in section 5.2. The test corpus originally contained 5,029 English words in 359 sentences in 32 files. However, we needed to reduce the amount of data for analysis for two reasons. Firstly, seven sentences were dropped when exporting the translation memory to a text file for unknown, possibly technical, reasons as this problem occurred for all nine post-editors. Secondly, one of the participant post-editors failed to post-edit one of the files, which contained 26 sentences. In order to ensure that we analyse the same data set for all post-editors, it was decided to omit this file from analysis. This resulted in a final data set, which contained 4,822 words in 326 sentences in 31 files.

Although the test data contained MT output and TM fuzzy matches, the main part of the quantitative analysis only considers post-editing of MT output, leaving the comparison of post-editing of MT output and editing of TM fuzzy matches to section 7.7. Table 7.1 summarises the number of sentences and words for MT output and TM fuzzy matches. All 31 files contain one or more MT sentences, while TM fuzzy matches appear in only 18 files.

	MT output	TM matches	Total
Sentences	269	57	326
Words	3,916	906	4,822

Table 7.1 Profile of analysed text

7.2 Summary statistics of PE speed and GTM

Table 7.2 summarises the average PE speed and the GTM score (explained in section 3.4.2) for each post-editor. ‘Coef. of SD’ columns show the standardised standard deviations, which makes it possible to directly compare the standard deviations between PE speed and GTM scores, that use different units of measurement.

	PE speed (words/min)				GTM score			
	Mean	Coef. of SD	Min	Max	Mean	Coef. of SD	Min	Max
Post-Editor A	16.68	81%	0.92	84.00	0.72	28%	0.17	1
Post-Editor B	17.39	204%	0.31	468.00	0.74	30%	0.14	1
Post-Editor C	24.33	83%	2.03	154.29	0.64	34%	0.00	1
Post-Editor D	17.81	91%	0.97	144.00	0.63	32%	0.00	1
Post-Editor E	20.24	98%	0.89	210.00	0.70	30%	0.00	1
Post-Editor F	19.93	137%	1.01	380.00	0.66	32%	0.14	1
Post-Editor G	23.87	204%	0.10	660.00	0.73	26%	0.00	1
Post-Editor H	23.11	86%	0.48	130.91	0.63	37%	0.00	1
Post-Editor I	38.04	86%	1.17	195.00	0.78	26%	0.00	1
Average of all	22.38	119%	0.88	269.58	0.69	30%	0.05	1

Table 7.2 Average PE speed and GTM scores

As can be seen from the table, the mean of PE speed is distributed within less than a ten second range, except for Post-Editor I, who has an exceptionally high speed and is

regarded as an outlier here.⁴³ The slowest PE speed of the remaining post-editors is for Post-Editor A at 16.68 words/min, which is the equivalent of approximately 1,000 words per hour, and the fastest, Post-Editor C, is 24.33 words/min, which is about 1,500 words per hour. These figures somewhat correspond with the figure Plitt & Masselot (2010) reported on post-editing of FIGS, translated from English by an SMT system; the throughput for their post-editors was 800 to 1,800 words per hour (see section 5.2.1.2).

GTM scores roughly show how much of the MT output was untouched during PE; for example, if the GTM score of a given sentence is 0.75, roughly 25% of the sentence was modified, and 75% was left intact. The distribution of the mean of GTM scores in Table 7.2 shows that some post-editors made changes on more than one third of the text, while others made changes on about one quarter of the text.

The values for standardised coefficient of standard deviation (Coef. of SD) tell us that both within- and between-post-editor variance is much higher for PE speed compared to GTM scores. This means that the amount of editing performed during PE is more or less similar within and between post-editors, while the time taken to make the changes varies greatly both within and between post-editors.

On the other hand, the fact that both PE speed and GTM scores show variance between post-editors indicates that some post-editors tend to make more changes than others, and some post-editors take more time than others. The question here is: If one makes more changes than others, is she/he slower to post-edit, and vice versa? In order to find out the answer to this question, we examined the correlation between the average GTM score (the amount of changes made) and the average PE speed for each post-editor. Figure 7.1 shows the results.

⁴³ We follow one of the common definitions of an outlier using the interquartile range (IQR): lower than the 25th quartile minus 1.5*IQR or greater than the 75th quartile plus 1.5*IQR. In this case, 25th quartile=17.81, 75th quartile=23.87, thus IQR=6.06. $23.87 + 1.5 * 6.06 < 38.04$.

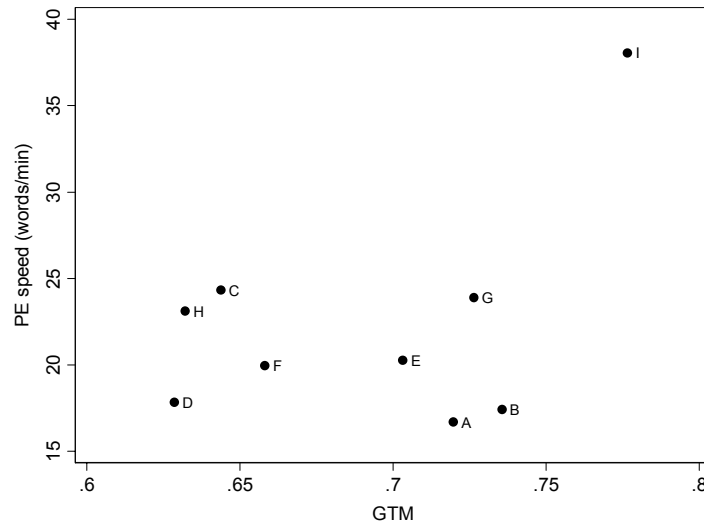


Figure 7.1 Correlation between average PE speed and GTM scores by post-editors

As can be seen, except for Post-Editor I, who is an outlier and shows an exceptionally high PE speed and relatively high GTM at the same time, there seems to be no clear relationship between the PE speed and the GTM score on a post-editor-by-post-editor basis. Therefore, it can be said that the post-editors that tend to make more changes than others are not necessarily slow post-editors, which supports the findings from similar studies (de Almeida & O’Brien 2010, Plitt & Masselot 2010).

7.3 Correlation between the amount of editing and the PE speed

As stated in section 3.1, our first research question is:

RQ1: How does the amount of editing correlate with the amount of effort in post-editing of English to Japanese MT output?

The following series of scatterplots show the correlation between the amount of editing performed during PE, measured by GTM, and the PE speed. A Pearson correlation coefficient for each post-editor is shown above each plot ($r=$). PE speed data have been transformed to logarithm numbers in order to secure normal distribution and a linear relationship with GTM scores for statistical analyses, in keeping with the method employed in the preliminary study as mentioned in section 4.3.1.

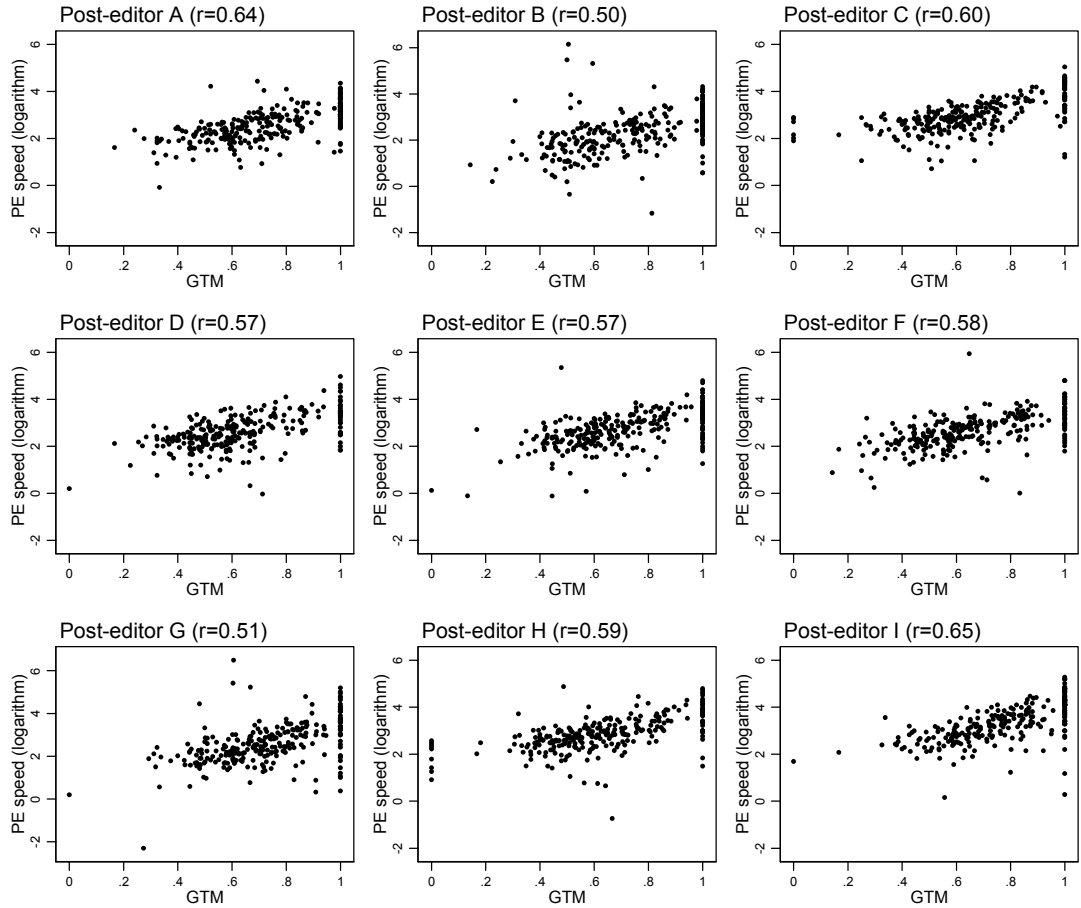


Figure 7.2 Correlation between PE speed and GTM scores for each sentence by post-editors

The Pearson correlation coefficients vary from 0.5 to 0.65 ($p < 0.001$, for all results), and the average over nine post-editors is 0.58, which is slightly higher than the result from the preliminary study (0.56), and proves a moderate correlation level, which somewhat contradicts Krings's findings mentioned in section 2.4.1. Krings's study suggests that the temporal PE effort of middle quality MT output, which had a middle level textual similarity with PE output, was larger than not only that of high quality MT (higher textual similarity), but also that of poor quality MT (lower textual similarity) (Krings 2001). However, the condition of the experiment in the present study differs greatly in a number of aspects from that of Krings. Firstly, MT systems have advanced significantly over the past two decades. Secondly, the tested language pairs are different. More importantly, the desired qualities of the final products are different; the post-editors were required to produce human translation level output in Krings's experiment, whereas in our experiment, they were asked to perform minimum post-editing in order

to ensure that the target text was correct and understandable, and any stylistic editing was discouraged.

In any case, the above figures, along with the results from the preliminary study, confirm that there is a proportional relationship between the amount of editing made during PE and the PE speed on a sentence-by-sentence basis. However, the variance is rather large, which may be explained by taking into account other factors, such as ST characteristics and PE operation types.

7.4 Effect of ST Characteristics

As discussed in Chapters 2 and 3, ST characteristics have been found to have impacts on the amount of PE effort. In the preliminary study, we considered two of such independent variables, namely, the structure and the length of the ST, and found that both of them had impacts on PE speed. In the present study, five additional variables are considered: component parts of the document, number of UI terms, number of technical terms, complexity index measured by the MT system, and conformity to controlled language rules measured by controlled language checking software. The following subsections explain the concepts behind these variables, and show the effect of them on PE speed separately.

7.4.1 ST length

As already discussed in section 4.4.2, sentence length often becomes an issue when using MT systems. Bernth & Gdaniec (2001) suggests that very long or short sentences should be avoided as they compromise ‘MTranslatability’. It has been suggested by some prior research that very long and very short sentences cause problems for MT systems (Underwood & Jongejan 2001) and thus require more PE effort (O’Brien 2006b, Aikawa et al. 2007). Very long sentences with multiple clauses can produce both grammatical and semantic complexity, while very short sentences, such as those with fewer than five words, may lack context and consequently be semantically ambiguous. The result from the preliminary study showed that the sentences of six to ten words in length are faster to post-edit than all other sentence lengths (Figure 4.5).

Figure 7.3 shows the findings from the present study. The box plot represents the distribution of average PE speed (words/min) of nine post-editors by ST length categories. The sentences were categorised into groups according to the number of words contained: 1-5, 6-10, 11-15, 16-20, 21-25, and over 25 words. The white line in each box shows the median value among nine post-editors, and the box represents the range of distribution in the interquartile range (IQR, the range between 25th and 75th percentile), which shows approximately the middle 50% of the data. The whiskers, the horizontal lines above and below each box, show the upper and the lower 1.5 IQR values above and below IQR respectively, and the dots represent the outliers. The number after ‘N=’ in parentheses under each length category indicates the number of observations found in the category.

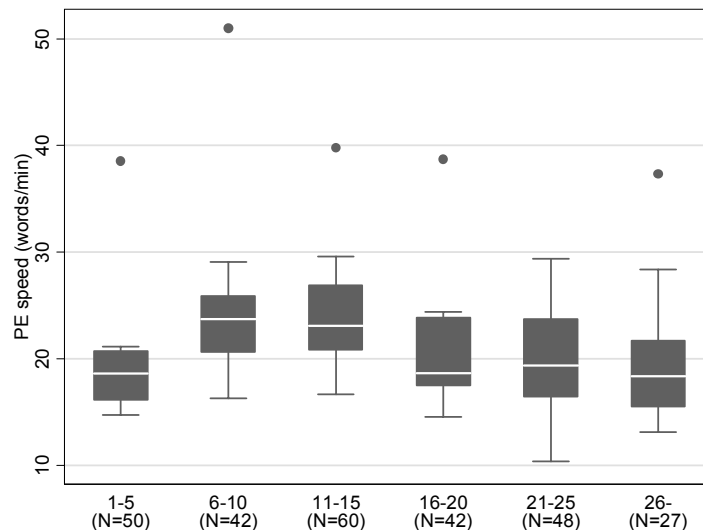


Figure 7.3 Average PE speed by ST length categories

According to the chart, both IQR and the median values suggest that the PE speed for the sentences that contain 1-5 words is comparatively slow, PE speed is fastest for the sentences of 6-15 words in length, and it gradually becomes slower again as the sentences become longer. This trend has been observed in the preliminary study (4.4.2) and also reported by Plitt & Masselot (2010), though in their study the optimum length for PE was found to be around 22 words. In any case, in order to take into consideration the diminishing effect of the ST length, it was decided that, in multiple regression analysis in section 7.6, we will employ the same method as we did in the preliminary

study, discussed in section 4.4.5.1, that is to use a square term along with the normal term of the variable.

7.4.2 ST structure

The sentence structure is one of the important aspects of controlled authoring. Kohl (2008) recommends to avoid using ‘interrupting sentences’, which partly corresponds to our definition of ‘incomplete sentences’ discussed in section 4.4.1, as they compromise clarity and translatability. He does not suggest that one should avoid using complex sentences, but gives several cautions when using subordinate clauses in a sentence, such as clarifying what a relative clause is modifying (ibid). In the preliminary study, we examined how the difference in sentence structures, namely, simple, complex, and incomplete sentences, affect the correlation between the PE speed and the amount of editing made during PE (section 4.4.1). In the present study, however, we will also check how the difference in structures affects the amount of PE effort measured by speed. Similar to the preliminary study, we had a rather small number of cases of compound sentences (N=11). We excluded the compound sentences from the analysis in the preliminary study. In this study, however, in order to secure as many sentences as possible to analyse, we would combine them with complex sentences and make them as one group since they both share the same fundamental characteristic of having multiple clauses.

The sentence structure has a close relationship with the sentence length. Table 7.3 shows the mean, minimum, and maximum sentence length and standard deviation of the mean for each sentence structure.

	Mean	Min	Max	SD
Simple	10.65	4	29	5.07
Complex/Compound	19.70	7	45	6.50
Incomplete	4.54	1	15	3.11

Table 7.3 Sentence length by sentence structure

On average, Simple sentences are longer than Incomplete sentences, and Complex/Compound sentences are longer than Simple sentences. However, the same does not apply to all sentences; as minimum and maximum values show, some

Complex/Compound sentences are shorter than some Simple or Incomplete sentences, some Incomplete sentences are longer than some Simple or Complex/Compound sentences, and so on. Therefore, the sentence length and the sentence structure are considered separately in order to see the effect of them individually.

Figure 7.4 shows the distribution of average PE speed from nine post-editors by sentence structure.

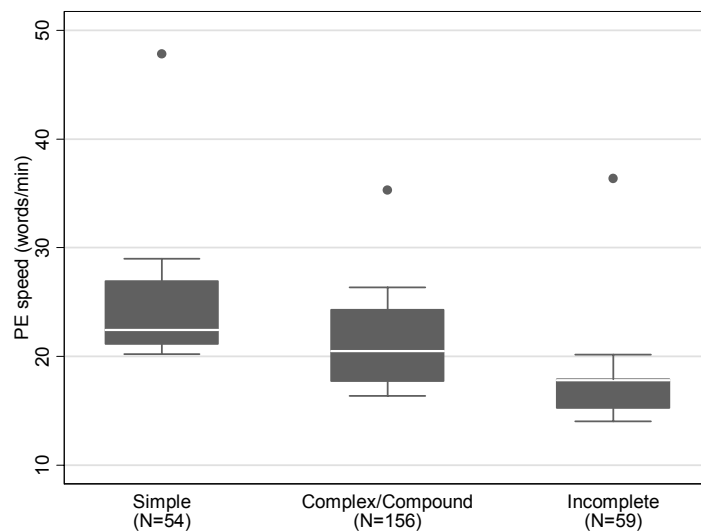


Figure 7.4 Average PE speed by ST structures

As can be seen, Simple sentences are the fastest to post-edit, Complex/Compound sentences come next, and Incomplete sentences are the slowest to post-edit. This is interesting especially when compared to the amount of editing for each sentence types, as shown in Table 7.4.

	GTM scores	PE speed (words/min)
Simple	0.74 (0.21)	25.92(25.54)
Complex/Compound	0.66 (0.18)	22.42(32.35)
Incomplete	0.73 (0.28)	19.02(19.11)

Standard deviations are shown in parentheses.

Table 7.4 Mean GTM scores and PE speed by sentence types

While the amount of editing, measured by GTM scores, is larger for Complex/Compound sentences compared to Incomplete sentences, the PE speed is faster for Complex/Compound sentences compared to Incomplete sentences.

In multiple regression analyses, we will check the effect of sentence types on PE speed with other independent variables fixed. In addition, it is worth noting that it can be observed from the box plot that the difference in the size of the box suggests that while the average PE speed for Simple and Complex/Compound sentences has a considerable variance among post-editors, the PE speed for Incomplete sentences is more uniformly slow for all post-editors.

7.4.3 Component parts of the document

One of the characteristics of technical documents, especially IT-related documents including software manuals, compared to other types of documents, such as novels and newspaper articles, is that they contain a number of different component parts. For example, they often contain bullet lists, numbered imperative sentences for explaining procedures, table cell entries, titles in various styles for paragraphs, figures, and tables, and so on. It was of interest whether some component parts are easier to post-edit than others. To which document component part the sentence belongs has been determined automatically based on the XML tags embedded in the original document, as mentioned in section 5.2.1.4. However, as Table 7.5 summarises, some parts, namely, Figure Title, List Title, Procedure Title, and Table Title, appear only a very few times in the entire data set, which makes these document parts unsuitable for statistical analysis. Therefore, we decided to aggregate these parts into one category: Various Titles.

Parts	Frequency
Section Title	25
General Sentence	119
List Title	4
List Item	21
Procedure Title	6
Procedure Step	64
Figure Title	1
Table Title	1
Table Cell Contents	28
Total	269

Table 7.5 Frequency of each document component part in the test corpus

Figure 7.5 summarises the distribution of average PE speed of the nine post-editors by document component parts. As can be seen, the Procedure Step category has the fastest

PE speed with a relatively broad distribution, while the List Item category has the slowest PE speed for all post-editors. Other parts that have relatively slow PE speed include Section Title and Table Cell.

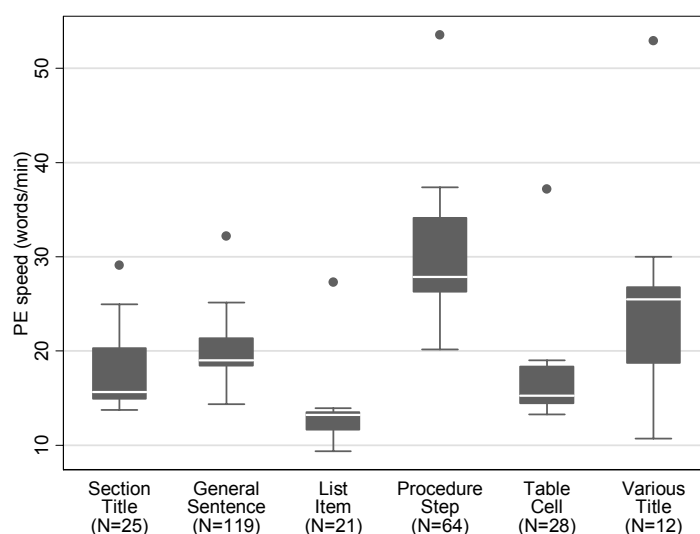


Figure 7.5 Average PE speed by document component parts

The reason for this may partly be explained by further examining the characteristics of each document part category. Table 7.6 shows the cross tabulation information of document parts and ST structures. The bold face shows the actual number of sentences in each category, and the figures in parentheses show their percentages.

Component Parts	Structure			Total
	Simple	Complex/Compound	Incomplete	
Section Title	2 (8%)	0 (0%)	23 (92%)	25 (100%)
General Sentence	25 (21%)	93 (78%)	1 (1%)	119 (100%)
List Item	7 (33%)	7 (34%)	7 (33%)	21 (100%)
Procedure Step	17 (27%)	46 (72%)	1 (2%)	64 (100%)
Table Cell Contents	1 (4%)	7 (25%)	20 (71%)	28 (100%)
Various Title	2 (17%)	3 (25%)	7 (58%)	12 (100%)
Total	54 (20%)	145 (54%)	59 (22%)	269 (100%)

Table 7.6 Cross tabulation of document component parts and ST structures

The Procedure Step category consists mainly of Complex/Compound and Simple sentences, with a very small proportion of Incomplete sentences, the sentences that are slowest to post-edit, as mentioned in section 7.4.2. This may be one of the reasons for the fastest PE speed for this category. However, this does not explain the disparity between the PE speed for Procedure Step and General Sentence, both of which have a

similar proportion of ST structures. This may partly be explained by the uniformity of the Procedure Step sentences, most of which follow certain patterns such as:

Imperative only:

Select one or more department.

Click <guimenuitem> Sign up for News Bulletins </guimenuitem> .

Imperative with a modifier:

In the folder home page, click Role Assignment.

Click one or more departments that you want to delete.

The Section Title and the Table Cell categories have higher numbers of Incomplete Sentences than other categories, which may partly explain the slow PE speed, but again, the category with the slowest PE speed, List Item, does not have a very high population of Incomplete Sentences; in fact, although one might assume ‘list items’ are not usually complete sentences, the XML tags that have determined the document parts are used for formatting purposes and not for representing linguistic characteristics, and List Item category in our test corpus contains the same number of simple, complex/compound, and incomplete sentences (Table 7.6). This may suggest that the two independent variables, ST structure and document component part, may interfere with each other to some extent, and do not correlate perfectly, thus it is worth including both these independent variables to see the combined effect. In the multiple regression analysis, we will take into account the Procedure Step category, the sentences with the highest PE speed, and the List Item category, the sentences with the lowest PE speed, to see if there is a notable difference in PE speed between these two document component parts.

7.4.4 Number of UI terms

As Roturier & Lehmann (2009) suggest, checking and correcting UI terms, such as menu names, option names, and messages that appear on the product interface, is a time consuming task for humans, as they need to be standardised to match company specific terminology. As discussed in section 5.2.3.1, we observed an interesting phenomenon in the preliminary study where the post-editors sometimes made changes to the machine

translated UI terms even though they were clearly instructed not to alter the translation of those terms even when they had been translated into obviously inappropriate Japanese words. The participant post-editors had no prior knowledge of Symantec's terminology, nor were they provided with any product-specific glossary, but it seemed to be difficult for them to ignore UI terms even though they had no means of checking the correctness of those terms. This raised a question as to whether the presence of UI terms distracted post-editors, and moreover, whether it would help to reduce this distraction if those terms were left in English. In fact, in some real-life situations, UI terms are left in English throughout the translation and PE phases, and standardised on the client's side after PE. In the present study, UI terms are clearly marked by <guimenuitem> XML tags. In order to compare the PE speed between the sentences that contain translated UI terms and those that contain non-translated UI terms, we included some translated UI terms and some non-translated UI terms. The original ST corpus included 60 UI terms, 29 of which were translated and 31 left in English in the machine-translated corpus. However, as explained in section 7.1, we had to drop some of the sentences from the original test corpus, which also resulted in a decreased number of UI terms; the remaining corpus contained 27 translated UI terms and 29 non-translated UI terms.

Figure 7.6 shows the distributions of the average PE speed of nine post-editors by the number of UI terms, both translated and non-translated. 228 sentences contained no UI terms, and the largest number of UI terms contained in a sentence was three.

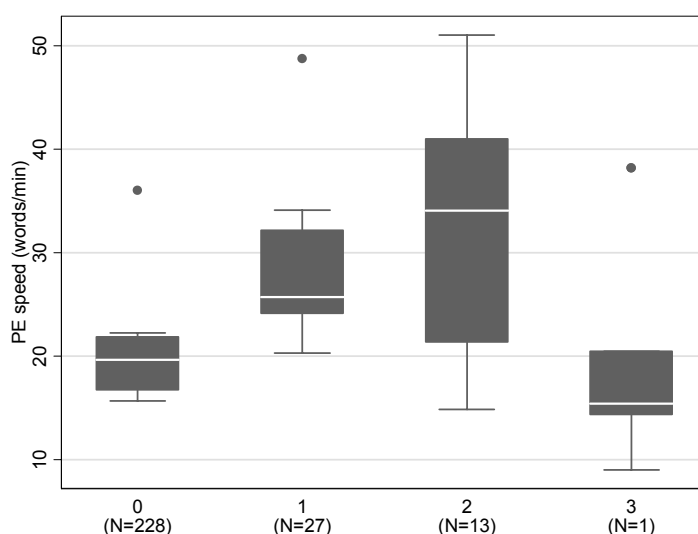


Figure 7.6 Average PE speed by the number of UI terms

Interestingly, the figure suggests that containing a larger number of UI terms in a sentence results in faster PE speed (excluding the single case that contains three UI terms), which contradicts our assumption that the presence of UI terms may slow down the PE process. However, this relationship might have been influenced by other factors, such as sentence structure and document component type. For example, 59 out of 228 sentences in the UI terms=0 group are Incomplete sentences (in fact, all Incomplete sentences belong to this group), which have been found to be slower to post-edit than other sentences (discussed in section 7.4.2). On the other hand, 78% of the sentences in the UI terms=1 group and all sentences in the UI terms=2 group are categorised in the document component type ‘Procedure Step’, which have been found to be faster to post-edit than other document component types (section 7.4.3). Therefore, we need to wait until we perform multiple regression analysis to decide the true effect of UI terms on PE speed independent of other conditions. In order to see the effect of translated and non-translated UI terms separately, we will employ two separate variables: one for translated UI terms and another for non-translated UI terms. Since the sentence that contains three UI terms shows clearly a different trend from others, which compromises the linearity, and since there is only one such case, we will omit this observation from multiple regression analysis.

7.4.5 Number of Technical terms

In the present study, we also consider the number of Technical terms, which are defined as either 1) the terms that are product specific but not marked as UI terms, or 2) other technical terms that are standardised within Symantec (and in some cases in the IT sector in general) and listed in Symantec’s glossary. Unlike UI terms, which are marked with XML tags, Technical terms appear as unmarked in the sentences. The distinction between UI and Technical terms here is in line with the one used for coding of PE operation types discussed in section 6.2.1.3. The criticalness of terminology consistency for this category is not as strict as for the UI terms, but it is still desirable to standardise these terms, thus it is suspected that post-editors may spend more time when technical terms are present in a sentence. Technical terms are all translated by the MT system. Figure 7.7 shows the distributions of the average PE speed of nine post-editors by the number of Technical terms in a sentence.

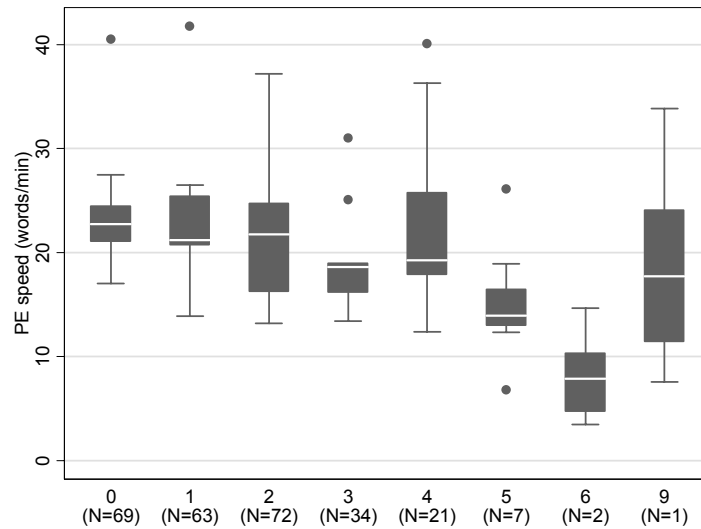


Figure 7.7 Average PE speed by the number of technical terms

The figure generally indicates that having more Technical terms in a sentence slows down the PE speed, except for the single case that contains nine Technical terms. We will check if this effect holds true when other conditions are also taken into account, by using multiple regression analyses. Similar to the case of UI terms, since the sentence that contains nine technical terms shows a different trend from others, and since there is only one such case, we will omit this observation from multiple regression analysis.

7.4.6 Complexity index by MT system

Systran version 6, the MT system we employed for the present study, offers a function that measures the syntactic complexity of the source sentences. It takes into account a number of aspects of the ST, including “the number of clauses, conjunctions, phrases in parentheses, prepositional phrases, sentence length, sentence type (question or declarative sentence) as well as multiple additional language-specific criteria” (SYSTRAN : p.141), and calculates the scores for each sentence; the lowest score is 1, and the higher it becomes the more complex the sentence is.⁴⁴ Systran version 6 provides this function to help the authors to produce simpler ST in order to improve the quality of the MT output. Therefore, we want to test the assumption that this index can be correlated with the amount of PE effort. Figure 7.8 shows the distributions of the average PE speed of nine post-editors by the Systran complexity scores.

⁴⁴ SYSTRAN’s documantation does not mention any possible maximum value.

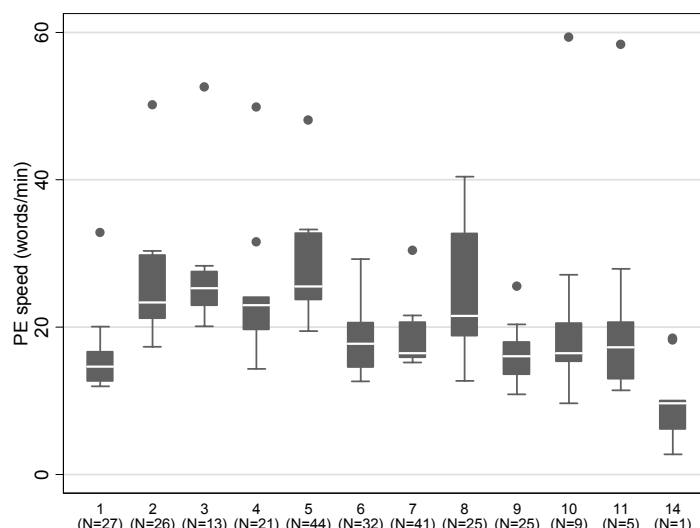


Figure 7.8 Average PE speed by complexity index scores

The figure shows that the sentences with complexity scores between 2 and 5 are generally faster to post-edit than sentences with scores between 6 and 14. It may appear surprising that the sentences with a complexity score of 1 are among the slower ones, but this may partly be explained by the fact that all 27 sentences with a score of 1 are incomplete sentences, which were indicated to be slower to post-edit than simple and complex sentences. It is also worth noting that, based on the examination on our test corpus, the Systran complexity index had a high correlation with ST length (Pearson correlation coefficient=0.89), even though it takes into account other aspects mentioned above. This may be one of the reasons that, although the figure suggests the general trend that the higher the score (more syntactic complexity), the slower the PE speed, the relationship is not perfectly linear. This will be further investigated by using multiple regression analysis.

7.4.7 Conformity to controlled language rules

As mentioned in section 2.4.2.1, the impact of CL rule conformity of the ST on the quality of MT output and the amount of resulting PE effort has been extensively investigated. In this study, however, we did not examine each sentence or CL rule qualitatively; instead, we employed the automatic measurement that gives an idea of the level of CL rule conformity of each sentence. The function is provided by CL checking

software, acrolinx IQ⁴⁵ (formerly ‘acrocheck’), for which all the Symantec CL rules are defined. Acrolinx IQ checks the ST in terms of grammar and stylistic appropriateness, and provides two types of measurements: *flags*, which indicate the absolute number of detected problems, and *scores*, which are the normalized values of flags in relation to the sentence length. We employed the flag count for two reasons: 1) it is more suitable for sentence level analyses; the scoring mechanism is designed for document level analysis,⁴⁶ and 2) scores were distributed in a heavily skewed manner, as 84% of the sentences were given a score of 0, and the remaining scores were scattered in the range between 250 and 5,000. Even though the distribution of flags was also skewed, it was less skewed compared to the scores as flag counts fell in the range between only zero and four. Figure 7.9 shows the distributions of the average PE speed of nine post-editors by the acrolinx IQ CL rules conformity flag counts.

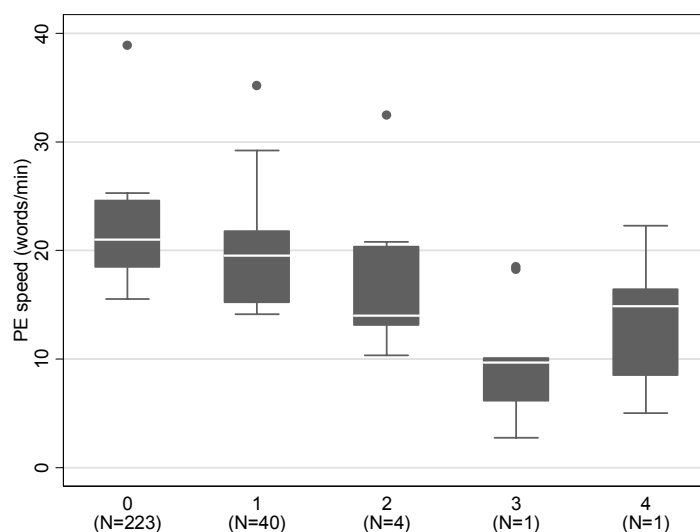


Figure 7.9 Average PE speed by CL rule conformity

The result shows that the sentences with a score of zero are faster to post-edit than the sentences that contain CL rule violations, and as the number of violations increases, the PE speed slows down, except for the sentence with a flag count of 4, for which only one case was observed. Overall, the figure suggests that the relationship between CL rule

⁴⁵ <http://www.acrolinx.com/> [Last accessed: 19/10/2010]

⁴⁶ <http://www.acrolinx.com/uploads/documents/doc-center/acrolinxIQSuite1.0/Plug-inUserGuides/EN/acrocheck%20for%20Word%20Plug-in%20User%20Guide.pdf> [Last accessed: 19/10/2010]

conformity level measured by acrolinx IQ and the PE speed is somewhat linear. This supports the general findings from relevant studies that have used different measurement methods and languages (O'Brien & Roturier 2007, O'Brien 2007). Therefore, it was decided to include this variable in multiple regression analysis. Again, as the sentence that has a score of 4 shows a different trend from others, and since there is only one such case, we will omit this observation from multiple regression analysis.

7.5 Effect of Post-Editing Operation Types

As discussed in Chapter 6, PE operation types were defined using a qualitative investigation of each edit, and assigned to each edited morph by the researcher. We first examined the frequency of edits in each category.

7.5.1 Edit frequency of each PE type

Table 7.7 shows the number of morphs categorised into each PE operation type.

	Sup	Oms	UI	Tech	Gen	B-Sty	Dep	Rwt	S-Sty	Punc
Post-Editor A	71	64	9	149	334	166	334	349	97	28
Post-Editor B	98	33	5	39	383	138	303	501	213	69
Post-Editor C	114	54	6	107	437	109	284	643	353	70
Post-Editor D	101	38	0	91	383	185	283	734	491	49
Post-Editor E	49	35	34	89	328	79	203	707	256	29
Post-Editor F	62	39	43	80	340	160	223	857	265	46
Post-Editor G	59	36	3	27	253	131	183	755	158	60
Post-Editor H	63	49	27	74	370	139	194	944	446	36
Post-Editor I	15	46	2	19	245	19	179	657	177	6
Average	70	44	14	75	341	125	243	683	273	44
(Std.Dev.)	(30)	(10)	(16)	(41)	(62)	(51)	(58)	(178)	(133)	(21)

Sup=Supplementation, Oms=Omission, UI=UI Term Alteration, Tech=Technical Term Alteration, Gen=General Term Alteration, B-Sty=*Bunsetsu* Level Stylistic Change, Dep=Dependency Edit, Rwt=Rewrite, S-Sty=Structure Level Stylistic Change, Punc=Punctuation

Table 7.7 Frequency of PE operation by post-editors

Naturally, the numbers of edits in the UI category are very small; the post-editors are not supposed to edit UI terms at all. Similarly, the relatively small number of edits in the B-Sty category may indicate that the post-editors followed the instruction in the PE guidelines that discourages stylistic edits, at least to some degree. Other general tendencies can be seen in the frequency of edits in each category, for instance, a large

number in the Rwt category and a smaller number in Dep, S-Sty, Gen, and even smaller in Oms and Punc. However, the rather large number in standard deviation tells us that the frequency varies across post-editors.

The definition of frequency in terms of PE operation is not straightforward in the first place, since one *bunsetsu*, the smallest unit of meaning in Japanese, can contain one or more morph(s), and thus the number of morphs contained in ‘one unit of PE operation’ varies. As the examples in Table 7.8 show, an edit often happens on multiple consecutive morphs. In fact, editing a single morph with surrounding morphs intact occurs in only less than 10% of edits; most of the edits are performed on multiple consecutive morphs.

MT Output	Part of Speech	PE Output	Part of Speech	Technical Effort	PE Operation
他	Noun	他	Noun		
の	P-case	の	P-case		
ロール	Noun-Sa	ロール	Noun-Sa		
		で	P-case	Ins	Dep
は	P-adv	は	P-adv		
の	P-case	、	Punc	Sub	Punc
エクスポート	Noun	ユーザ	Noun	Sub	Rwt
の	P-case	が	P-case	Sub	Rwt
フォルダ	Noun	フォルダ	Noun	Shift: 3	Rwt
から	P-case	から	P-case	Shift: 3	Rwt
アイテム	Noun	アイテム	Noun		
に	P-case			Del	Rwt
ユーザー	Noun			Del	Rwt
を	P-case	を	P-case		
可能	Noun-adj	エクスポート	Noun	Sub	Rwt
し	Sa-Verb	し	Sa-Verb	Shift: -1	Rwt
に	P-case	たり	P-case	Sub	Rwt
、	Punc	、	Punc		
それ	Pron	フォルダ	Noun	Sub	Exp
に	P-case	に	P-case		
追加	Noun-Sa	追加	Noun-Sa		
する	Sa-Verb	する	Sa-Verb		
ため	Noun-dep			Del	Dep
に	P-case			Del	Dep
新しい	Adj	新しい	Adj		
アイテム	Noun	アイテム	Noun		
を	P-case			Del	Elp
有無	Noun			Del	Elp
を	P-case	を	P-case		
検索	Noun-Sa	検索	Noun-Sa		
し	Sa-Verb	でき	Verb	Sub	Gen
ます	Aux-verb	ます	Aux-verb		
。	Punc	。	Punc		

The number next to 'Shift' in the Technical Effort column indicates the distance of the move; 3=moved forward for 3 morphs, -1=moved backwards for 1 morph

Table 7.8 Example of PE operation classification

This makes it difficult to answer the question ‘How much each edit in each PE operation type category slows down the PE?’, as the definition of ‘each edit’ is not unequivocal. This needs to be taken into account when we interpret the substantive significance of the effect from multiple regression analysis, as the number of morphs involved in each edit category was used as the value of each variable.

7.5.2 Number of visits

In addition to the PE operation types, we also considered the number of ‘visits’ to each sentence during the PE task. A visit is defined as a sequence of events in which the post-editor opens the translation segment, makes one or more changes or simply reads the translation without making any changes, and closes the segment. In the macro discussed in section 5.1.1.2, we embedded a function that keeps record of how many times the segments have been freshly saved, so that we could obtain the information about the number of visits paid to each segment. The post-editors were instructed to visit every sentence at least once, even if no editing was necessary, to count the ‘reading’ time into the PE duration, thus the minimum number of visits is 1. Although revisiting the sentences for necessary correction was allowed, the post-editors were encouraged to avoid multiple visits as much as possible. Figure 7.10 shows how many sentences have been visited how many times by each post-editor.

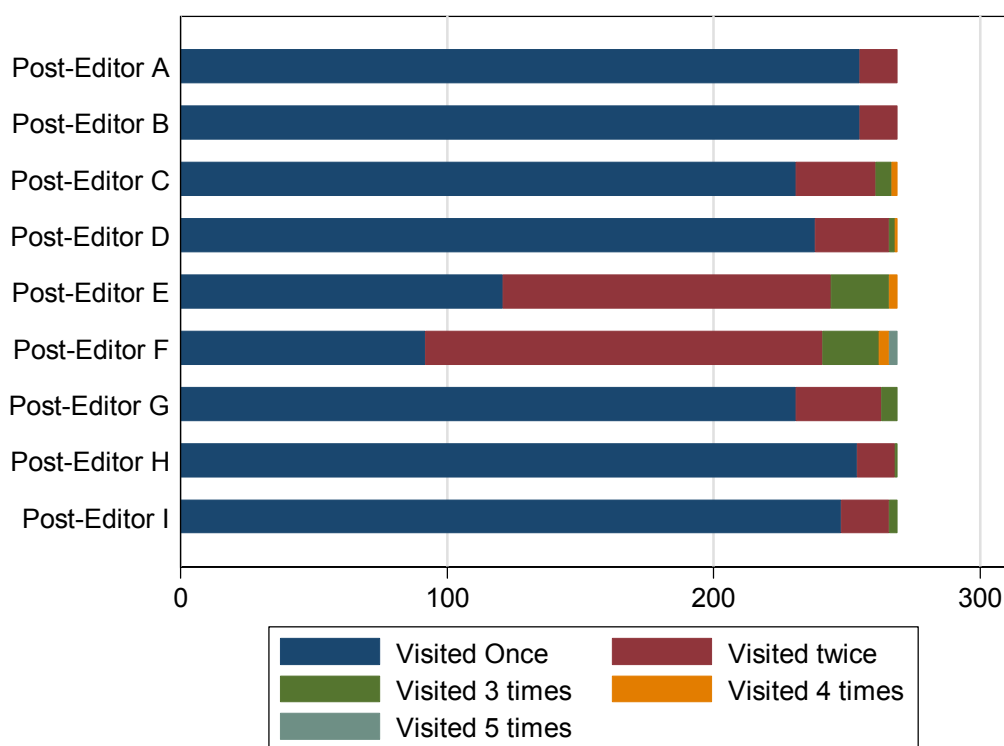


Figure 7.10 The number of sentences visited different times

As can be seen, revisits were generally avoided as instructed, except for Post-Editors E and F. It may be interesting to point out the resemblance of the ratio between Post-Editors A and B, C and D, E and F, and H and I. Each of these pairs was from the same

vendor, and participated in the experiment together at the same venue and time. This suggests that the revisits are not solely made from necessity, but may also be influenced by some environmental factors, such as PE training and quality control criteria of each vendor. Discovering the fact that the number of revisits greatly differ depending on post-editors raises a question about whether the post-editors that make more revisits are slower to post-edit. Table 7.9 compares the average number of visits and the average PE speed between post-editors. Figure 7.11 is a visual representation of the same data, ordered by highest average number of visits descending to lowest, accompanied by the average PE speed.

Post-Editor	Average PE speed (words/min)	Average # of visits
Post-Editor A	16.68	1.05
Post-Editor B	17.39	1.05
Post-Editor C	24.33	1.18
Post-Editor D	17.81	1.13
Post-Editor E	20.24	1.65
Post-Editor F	19.93	1.80
Post-Editor G	23.87	1.16
Post-Editor H	23.11	1.06
Post-Editor I	38.04	1.09
Average	22.38	1.24

Table 7.9 Average number of visits and PE speed

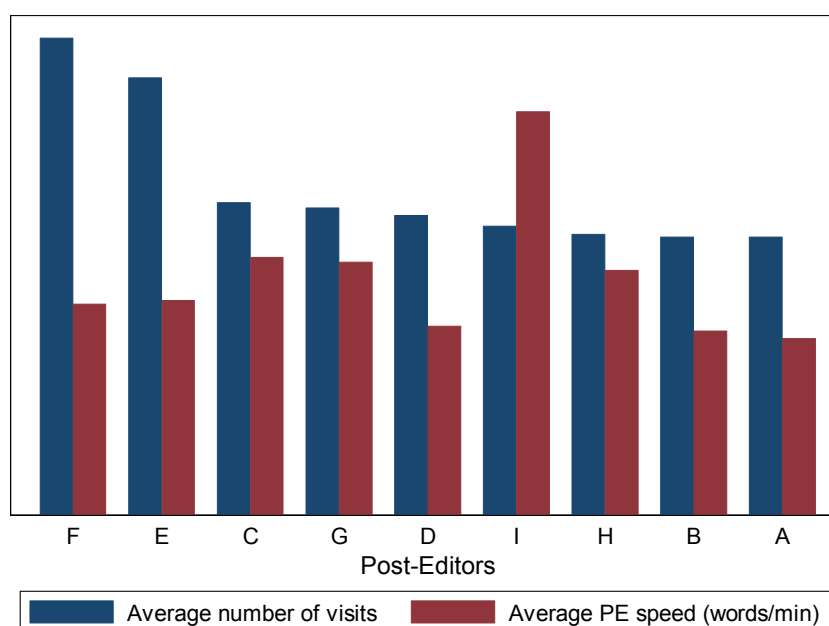


Figure 7.11 Comparison between the average number of visits and PE speed

According to the table and the figure, it is not true that making more visits makes one a slower post-editor. In fact, post-editors E and F, who have revisited more than half of the sentences, achieved a speed close to the average, especially to the average excluding Post-Editor I, the outlier here.⁴⁷ This ‘between-post-editor’ difference deserves more explanation, which will be investigated and discussed further and qualitatively in Chapter 8. In this chapter, however, we will focus on the quantitative aspect of revisits, and see if revisiting makes any ‘within-post-editor’ difference to PE speed. Figure 7.12 compares the distribution of PE speed for each sentence to the number of visits.

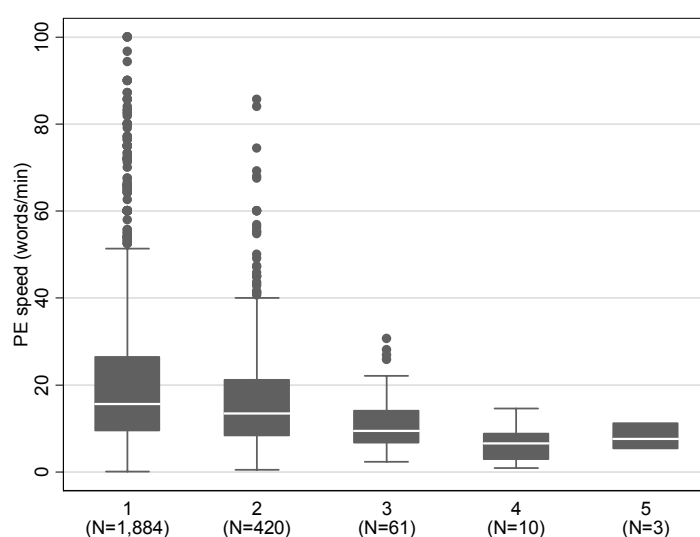


Figure 7.12 PE speed by number of visits

As can be seen, there is a relationship between PE speed and the number of visits, except for the sentences visited five times, of which there are only three cases. The figure suggests that there is a negative correlation between the PE speed and the number of visits when we look at the ‘within-post-editor’ effect. This will be further examined by multiple regression analyses in the next section. The segments that have been visited five times, however, will be omitted from multiple regression analysis since it compromises the linearity.

⁴⁷ The average excluding Post-Editor I: 20.42 words/min.

7.6 Effect of Combined Variables - Multiple Regression Analysis

In this section, the independent variables mentioned above will be tested by means of multiple regression analysis to see the effect of each independent variable on the PE speed holding other conditions fixed, and also the collective effect of various independent variables. Table 7.10 shows the results from multiple regression analyses based on different models. Model I tests all above-mentioned independent variables together in all sentences, while Model II is a modified version where only highly statistically significant independent variables ($p < 0.01$) are considered. Among tested independent variables in Model I, Technical Term and CL Rules Conformity Flags from the Source Text Characteristics category do not have enough statistical significance to be proven as effective, so were omitted in the refined models. Likewise, Omission, UI Term, Technical Term, *Bunsetsu* Style, and Structure Style from the PE Operation category were also omitted due to a lack of statistical significance. In addition, Post-Editor C and Post-Editor H are omitted for the same reason. Models III, IV, and V are tested on sentences of three different structure categories respectively, and thus differ in the number of cases. As explained in section 7.3, the logarithm of PE speed, instead of the words-per-minute speed, is used as the dependent variable to secure the linearity necessary for this statistical analysis. Unstandardised coefficients, which are shown in bold face, indicate the *substantive significance* of the effect of each independent variable; a positive value indicates a positive correlation with the PE speed (the higher the value, the faster the PE speed), and a negative value indicates a negative correlation with the PE speed. The size of an impact, however, depends on the unit used for each variable, thus interpreting a substantive significance requires calculation, which will be shown in a footnote for each instance. Asterisks represent the *statistical significance* (**: $p < 0.01$, ***: $p < 0.001$). The *standard error* is shown in parenthesis under each coefficient. In addition to the independent variables discussed earlier in this chapter, we employed binary variables to cancel out the differences in general PE speed between post-editors. The coefficients for Post-Editor A, B, C, D, E, F, G, and H represent the speed difference in relation to Post-Editor I. R-squared is a measurement of goodness-of-fit of the regression model, and shows the ratio of the variation the independent variables in the model can collectively explain to the total variation in the data. It ranges from zero to 1.

Model:	I	II	III	IV	V
Sample:	All sentences	All sentences	Simple sentences	Complex/Compound sentences	Incomplete sentences
Primary Independent Variable:					
GTM Score (Range: 0 - 1)	1.812 *** (0.11)	1.944 *** (0.08)	2.181 *** (0.19)	2.220 *** (0.13)	1.548 *** (0.14)
Source Text Characteristics:					
ST Length (Range: 1-45)	0.057 *** (0.01)	0.054 *** (0.01)	0.044 (0.03)	-0.010 (0.01)	0.224 *** (0.04)
ST Length^2 (Range: 1-45)	-0.001 ** (0.00)	-0.001 *** (0.00)	0.001 (0.00)	0.001 (0.00)	-0.007 ** (0.00)
Complex/Compound Sentence (0 or 1)	-0.150 *** (0.05)	-0.143 *** (0.04)			
Incomplete Sentence (0 or 1)	-0.207 *** (0.05)	-0.217 *** (0.05)			
List Item (0 or 1)	-0.177 *** (0.06)	-0.176 ** (0.05)	-0.088 (0.08)	0.103 (0.06)	-0.210 (0.14)
Procedure Step (0 or 1)	0.259 *** (0.03)	0.254 *** (0.04)	0.318 *** (0.07)	0.198 *** (0.03)	-0.315 (0.27)
UI: non-translated (Range: 0 - 2)	-0.120 ** (0.04)	-0.088 (0.04)	0.170 (0.08)	-0.953 (0.05)	n/a
UI: translated (Range: 0 - 2)	-0.260 *** (0.04)	-0.210 *** (0.04)	-0.220 ** (0.09)	-0.118 *** (0.05)	n/a
Technical Term (Range: 0-8)	-0.028 (0.01)				
Complexity Index Score (Range: 1 - 14)	-0.032 ** (0.01)	-0.030 ** (0.01)	-0.048 (0.04)	-0.034 ** (0.01)	-0.073 (0.03)
CL Rules Conformity Flag (Range: 0 - 4)	-0.059 (0.03)				
PE Operation:					
Supplementation (Range: 0 - 8)	-0.073 *** (0.02)	-0.068 *** (0.02)	-0.012 (0.03)	-0.043 ** (0.02)	-0.256 *** (0.06)
Omission (Range: 0 - 7)	0.000 (0.02)				
UI Term (Range: 0 - 17)	0.020 (0.02)				
Technical Term (Range: 0 - 12)	-0.017 (0.01)				
General Term (Range: 0 - 21)	-0.025 *** (0.01)	-0.018 ** (0.01)	-0.025 (0.02)	-0.007 (0.01)	-0.033 (0.03)
<i>Bunsetsu</i> Style (Range: 0 - 12)	-0.010 (0.01)				
Dependency (Range: 0 - 17)	-0.024 ** (0.01)	-0.018 ** (0.01)	-0.001 (0.03)	-0.013 (0.01)	-0.068 ** (0.02)
Rewrite	-0.021 ***	-0.017 ***	-0.015	-0.013 ***	-0.046 **

(Range: 0 - 40)	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)
Structure Style (Range: 0 - 12)	-0.016 (0.01)				
Punctuation (Range: 0 - 3)	-0.133 *** (0.03)	-0.136 *** (0.03)	-0.114 (0.06)	-0.092 ** (0.04)	-1.167 (0.69)
Number of Visits (Range: 1 - 4)	-0.272 *** (0.03)	-0.264 *** (0.03)	-0.381 *** (0.07)	-0.147 *** (0.04)	-0.441 *** (0.07)
Post-Editor Variance:					
Post-Editor A (0 or 1)	-0.639 *** (0.05)	-0.592 *** (0.05)	-0.573 *** (0.11)	-0.636 *** (0.05)	-0.551 *** (0.09)
Post-Editor B (0 or 1)	-0.861 *** (0.05)	-0.817 *** (0.05)	-0.815 *** (0.11)	-0.807 *** (0.07)	-0.925 *** (0.14)
Post-Editor C (0 or 1)	-0.011 (0.05)				
Post-Editor D (0 or 1)	-0.349 *** (0.06)	-0.318 *** (0.05)	-0.240 ** (0.11)	-0.307 *** (0.05)	-0.420 *** (0.10)
Post-Editor E (0 or 1)	-0.287 *** (0.06)	-0.246 *** (0.06)	-0.153 (0.11)	-0.271 *** (0.05)	-0.310 ** (0.10)
Post-Editor F (0 or 1)	-0.176 ** (0.06)	-0.134 ** (0.06)	-0.000 (0.11)	-0.197 *** (0.05)	-0.098 (0.11)
Post-Editor G (0 or 1)	-0.583 *** (0.05)	-0.534 *** (0.05)	-0.399 *** (0.11)	-0.523 *** (0.06)	-0.700 *** (0.15)
Post-Editor H (0 or 1)	-0.114 (0.05)				
R-squared	0.49	0.48	0.54	0.50	0.48
Number of cases	2,391	2,391	486	1,375	530
Dependent variable: Logarithm of post-editing speed (Range: -2.303 to 6.492, SD: 0.853) Unstandardised coefficients are shown in bold face, with standard errors in the parenthesis underneath. Asterisks indicate statistical significance. **: p < 0.01, ***: p < 0.001					

Table 7.10 Regression analysis results of the effect on PE speed

7.6.1 Primary Independent Variable category

GTM Score, our primary independent variable, proves statistically significant for all models (shown by asterisks). The coefficients shown in bold face, substantive significance, indicate how much GTM scores correlate with the PE speed holding other conditions fixed. The coefficient in Model II, 1.944, suggests that the PE speed for GTM=1 sentences (meaning no editing has been performed and only read and confirmed as correct by the post-editor) is approximately 6 times faster⁴⁸ compared to GTM=0 sentences (meaning it needed to be entirely rewritten). We can see a more

⁴⁸ The coefficient calculated from a logarithmic number can only be interpreted as a multiplier or a percentage, instead of 'words/min' speed unit. The applicable equation is: $\exp(\text{coefficient}) - 1$. In this case, $\exp(1.944) - 1 = 5.98$.

granular effect by breaking down the GTM scores into units of 0.1, in which case the coefficient becomes 0.1944. An increase of 0.1 points in the GTM score appears to increase the PE speed by approximately 21%.⁴⁹ However, the effect differs greatly depending on the sentence structures. The coefficients for GTM in Models III, IV, and V suggest that the same increase speeds up PE by 24% and 25% for Simple and Complex/Compound sentences, respectively, but only by 17% for Incomplete sentences.⁵⁰ This suggests that, for Incomplete sentences, factors other than the amount of editing have a larger impact on PE speed than for Simple and Complex/Compound sentences. In fact, the coefficients for all the variables other than GTM Score listed under Model V are higher than those for Models III and IV.

7.6.2 Source Text Characteristics category

The unstandardised coefficients in this category represents the size of the effect each ST element has on increasing or decreasing the PE speed. **ST Length** and the square term of ST length (**ST Length²**) both have statistical significance for All sentences and Incomplete sentences, but not for Simple and Complex/Compound sentences. The substantive significance is much greater for Incomplete sentences than for other sentences, which corresponds to the findings from the preliminary study. However, the substantive significance is generally weaker compared to the results from the preliminary study. We used only a small set of independent variables in the preliminary study, whereas we took into account many more independent variables in the present study, thus we suspect that the effect of the sentence length observed in the preliminary study and section 7.4.1 should actually be attributed to other factors we took into account in this regression analysis.

The negative values of coefficients for **Complex/Compound Sentence** and **Incomplete Sentence** in Model II suggest that the PE speed for these sentences is slower than that of simple sentences. The figures confirm the findings in section 7.4.2. In general, complex/compound sentences and incomplete sentences are 13% and 20% slower to

⁴⁹ $100 * (\exp(0.1944) - 1) = 21.4$.

⁵⁰ $100 * (\exp(0.2181) - 1) = 24.3$, $100 * (\exp(0.2220) - 1) = 24.9$, $100 * (\exp(0.1548) - 1) = 16.7$

post-edit respectively⁵¹ compared to simple sentences even after considering other conditions.

The document component parts **List Item** shows statistical significance when considering all sentences, suggesting that the List Item sentences are about 16% slower than other document component parts⁵², but it loses its significance when looking at different sentence structures separately. On the other hand, **Procedure Step** constantly proves to be faster to post-edit compared to other document component parts, except for incomplete sentences where there is only one such case. The evidence is especially strong for simple sentences, which suggests that Procedure Step is about 37% faster to post-edit than other document component parts when only considering simple sentences.⁵³ For complex/compound sentences, the PE speed is shown to be 22% faster.⁵⁴

The presence of **UI** terms, whether in English or translated, proved to slow down the PE when considering all sentences, and this is even more the case when the UI terms have been machine-translated, especially for Simple and Complex/Compound sentences (none of the Incomplete sentences in the test corpus contain UI terms). In general, the presence of one translated UI terms in a sentence slows down PE by 20% and 11% respectively,⁵⁵ and this is regardless of whether or not the UI term is altered by the post-editor. This partly supports our assumption, discussed in sections 5.2.3.1 and 7.4.4, that the post-editors may be distracted by the existence of UI terms, but leaving them in English may help to reduce the distraction.

The coefficients for **Complexity Index Score** have statistical significance especially for complex/compound sentences, suggesting this score may help to predict the amount of PE effort to some extent. The range of this score spanned from 1 to 14 for our data, but about 95% of the sentences fell under the score 10. According to our estimate, if the scores of two complex/compound sentences are different by five points, the sentence

⁵¹ $100 * (\exp(-0.143) - 1) = -13.3$, $100 * (\exp(-0.217) - 1) = -19.5$

⁵² $100 * (\exp(-0.176) - 1) = -16.1$

⁵³ $100 * (\exp(0.318) - 1) = 37.4$

⁵⁴ $100 * (\exp(0.198) - 1) = 21.9$

⁵⁵ $100 * (\exp(-0.220) - 1) = -19.7$, $100 * (\exp(-0.118) - 1) = -11.1$

with a higher score can be predicted to be approximately 16% slower to post-edit than the sentence with a lower score.⁵⁶ On the other hand, CL Rules Conformity Flag did not prove statistically significant.

7.6.3 PE Operation category

The unstandardised coefficients in this category indicate each PE operation type's effort intensiveness. Among ten PE Operation types, five of them had statistical significance, namely, Supplementation, General Term Alteration, Dependency, Rewrite, and Punctuation. However, General Term Alteration and Dependency, which have rather low coefficients, lose statistical significance when looking at different sentence structures separately. A possible reason that the statistical significance for dependency edit is much lower compared to the result from the preliminary study is that, in the present study, when multiple dependency edits occur in one phrase/clause/sentence, it often led to a 'rewrite' instead of being counted as dependency edits. The reason that the substantive significance (the unstandardised coefficient values) are also much lower compared to the result from the preliminary study may be because we used the actual count of the dependency edit in the present study, as opposed to the binary variable used in the preliminary study, in which case the effect is summed up when multiple dependency edits occur in one sentence. **Supplementation** has a rather high coefficient for incomplete sentences, but more than 70% of the supplementations occur among complex/compound sentences. Since the average number of morphs for a single case of Supplementation is about 2, the coefficient of Supplementation for a complex/compound sentence suggests that each case of supplementation in complex/compound sentences slows down PE by approximately 8%.⁵⁷ Likewise, more than 75% of the **Rewrites** occur among complex/compound sentences, and the average number of morphs for a single case of rewrite is about 8. Therefore the coefficient of Rewrite for complex/compound sentences suggests that each case of rewrite in complex/compound sentences slows down the PE by approximately 10%.⁵⁸ **Punctuation** also has a high statistical significance for complex/compound sentences, and the coefficient is very high compared to other factors in PE operation categories.

⁵⁶ $100 * (\exp(-0.034 * 5) - 1) = -15.6$

⁵⁷ $100 * (\exp(-0.043 * 2) - 1) = -8.2$

⁵⁸ $100 * (\exp(-0.013 * 8) - 1) = -9.9$

Since punctuation edits mostly occur on a single morph, the figure suggests that a punctuation edit in complex/compound sentences slows down PE by approximately 9%.⁵⁹ This may appear to be rather high an impact considering the amount of text involved in a single case of punctuation edit, which is usually only one character. However, altering them can lead to a structural change of the sentence (Bernth & Gdaniec 2001) that affects not only naturalness of the language but sometimes meanings, as punctuation plays an important role as syntactic cues (Kohl 2008). This partly explains why punctuation edits could be one of the effort-intensive PE tasks. In relation to punctuation issues, O'Brien & Roturier (2007), in reporting the effect of different CL rules on PE effort on English to German translation, examined the impact of punctuation errors. They found that the incorrect use of semicolons in the ST had a high impact on PE effort, though that did not apply to other punctuations. In our data, however, there was no incorrect use of semicolon, and it cannot be proven that the punctuation problems have been caused by the English ST problem, thus the relationship between the two findings are not known. In any case, the result of our study shows that punctuation edits are one of the most effort intensive tasks among all PE operation types considered. This partly supports Temnikova's (2010) ranking of PE effort intensiveness, discussed in section 2.4.3, that considers punctuation-related edits as one of the most effort-intensive tasks.

Number of Visits retains statistical significance and a rather high substantive significance throughout all models, though the coefficient for complex/compound sentences is noticeably low compared to simple and incomplete sentences. Although the Number of Visits ranges from 1 to 5, more than 99% of the sentences have been visited less than four times.⁶⁰ If we take the coefficient for All sentences, one additional visit on the same sentence slows down the entire PE speed of the given sentence by 23%, and two additional visits slows it down by 41%.⁶¹ The issue of revisits will be further investigated qualitatively in Chapter 8.

⁵⁹ $100 * (\exp(-0.092) - 1) = -8.8$

⁶⁰ 1 visit=79%, 2 visits=17%, and 3 visits=3%

⁶¹ $100 * (\exp(-0.264) - 1) = -23.2$, $100 * (\exp(-0.264 * 2) - 1) = -41.0$

7.6.4 Post-Editor Variance category

Both statistical significance and substantive significance for most of the Post-Editor Variance is high, compared to other independent variables considered above, suggesting that post-editor variance has a high impact on overall PE speed. The differences in speed will be further investigated in Chapter 8.

7.6.5 Discussion

Some of the results from MLR appear to contradict the findings from single-variate analyses performed in section 7.4. For example, the presence of more UI terms proved to slow down PE speed in MLR, while it did otherwise in section 7.4.4. As discussed in section 7.4.4, this is probably because that when we consider only a single variable, the number of UI terms, we overlook the hidden effect of other factors, such as sentence structure and document component types. By using MLR and having other conditions fixed, we could see the actual effect of UI terms, that is, how the number of UI terms affect PE speed if other conditions, such as sentence length, structure, document component type, complexity, and so on, are controlled for. Similarly, the number of CL rule violations showed a clear negative relationship with PE speed in section 7.4.7, but did not have statistical significance in MLR. This is probably because that the CL Rules Conformity Flag correlates with other variables that have statistical significance, such as ST Length and Complexity Index Score (Pearson correlation coefficients=0.45 and 0.38 respectively, with $p<0.001$ for both). Since these two variables had stronger relationship with PE speed, the actual effect of CL Rules Conformity Flag on PE speed was proved to be nominal. These examples demonstrate the benefit of MLR, which can prevent us from being misled by the surface relationship observed by taking into account only a single variable, and help to reveal the true effect of each variable.

The overall results of MLR in Table 7.10 suggest that there are a number of independent variables that hold statistical significance when considering all sentences in the test corpus. However, many of them lose statistical significance when considering different sentence structures separately. GTM Score, Number of Visits, and some of the Post-Editor Variance, namely, those for Post-Editor A, B, D, and G are the only independent variables that hold statistical significance for all models. This demonstrates how difficult it is to identify common factors that lead to an increase in the amount of

PE effort and that can apply to all texts. This reflects O'Brien's conclusion that it is difficult to find common causes of PE effort across all text, and the individual variance among post-editors is high (O'Brien 2006b). R-squared values show that the above models can explain a little more than 50% of the variance at most, which may not be a particularly high ratio. However, Wooldridge (2006) warns that we should not focus too much on R-squared as it does not directly indicate the validity of the relationship between the dependent variable and each independent variable. The validity of the regression analysis was tested and will be discussed in Appendix C.

7.7 Comparison with TM match segments

In this section, we will examine the difference between post-editing of MT output and editing of TM fuzzy match segments from various points of view.

7.7.1 TM match ratio that corresponds to MT output

First of all, we compare the average editing speed (temporal effort) to see if there is a particular TM match threshold under which editing becomes more effort intensive than post-editing of MT output. As stated in section 7.1, the test corpus of the present study contains mostly MT output and a much smaller number of TM fuzzy matches (81% and 19% in terms of word count, respectively), thus may not be enough to make a statistical inference. Nonetheless, we attempted to make a comparison of effort between post-editing of MT output and editing of TM match segments of various match ratios by individual post-editors. Table 7.11 summarises the results of the comparison of PE speed.

	MT output	TM output				
	(N=269)	75-79% match (N=9)	80-84% match (N=14)	85-89% match (N=10)	90-94% match (N=12)	95-99% match (N=12)
Post-Editor A	16.68	8.55	13.35	14.56	17.27	20.61
Post-Editor B	17.39	12.73	11.01	14.51	45.02	15.71
Post-Editor C	24.33	24.09	34.26	32.80	37.70	56.57
Post-Editor D	17.81	14.57	21.54	25.64	37.55	30.61
Post-Editor E	20.24	14.18	18.41	23.35	31.82	28.83
Post-Editor F	19.93	14.02	24.07	20.55	22.47	18.24
Post-Editor G	23.87	14.33	16.07	37.33	40.13	32.75
Post-Editor H	23.11	22.20	25.55	31.13	40.84	34.69
Post-Editor I	38.04	36.57	31.66	40.20	53.45	47.79

Table 7.11 Comparison of average editing speed between MT output and TM matches

The average editing speed (words/min) of each TM fuzzy match category was compared with the average PE speed of MT output, and the closest figure is highlighted. A trend can be observed that the PE speed of MT output is closer to the lower TM fuzzy matches ('lower' not going below 75% in this case) than higher matches in general, but it spreads over all TM match ranges between 75 and 99% depending on the post-editors, and no decisive association can be found from the table. Overall, however, the average PE speed for MT output is *at least* faster than the average editing speed for 75-79% matches for all participants. In short, the PE speed of MT output across all post-editors lies within the range of editing speed for TM fuzzy matches above 75%, which means that the PE speed is not substantially lower than TM editing speed.

Now we turn our attention to the amount of editing (technical effort) and make a comparison between post-editing of MT output and editing of TM fuzzy matches, whose results are shown in Table 7.12.

	MT output	TM output				
	(N=269)	75-79% match (N=9)	80-84% match (N=14)	85-89% match (N=10)	90-94% match (N=12)	95-99% match (N=12)
Post-Editor A	0.72	0.63	0.77	0.84	0.86	0.86
Post-Editor B	0.74	0.73	0.80	0.87	0.88	0.90
Post-Editor C	0.64	0.75	0.79	0.77	0.87	0.88
Post-Editor D	0.63	0.67	0.80	0.85	0.85	0.85
Post-Editor E	0.70	0.63	0.74	0.80	0.84	0.88
Post-Editor F	0.66	0.63	0.77	0.78	0.78	0.74
Post-Editor G	0.73	0.74	0.76	0.92	0.91	0.86
Post-Editor H	0.63	0.67	0.75	0.84	0.90	0.89
Post-Editor I	0.78	0.75	0.79	0.90	0.92	0.91

Table 7.12 Comparison of average GTM scores between MT output and TM matches

Again, the figures from TM output that are closest to that of MT output are highlighted. As can be immediately seen, the average GTM scores for post-editing of MT output are closer to the lower TM fuzzy matches compared to the PE speed shown in Table 7.11. In fact, quite dissimilarly to the findings from the comparison of PE speed, the average GTM scores for MT output are lower than those of 75-79% fuzzy matches for four out of nine post-editors. This suggests that the amount of editing necessary for MT output tends to be larger than TM fuzzy matches above 75%, but the time taken to make those changes is shorter for MT output than TM fuzzy matches above 75%.

Guerberof (2008) compared the temporal effort in editing English to Spanish translation, and concluded that post-editing of MT output was faster than editing of 80-90% fuzzy match TM segments on average. This finding holds true for only two out of nine participants in our results (Table 7.11). Yamada (2010), on the other hand, compared the technical effort in editing English to Japanese translation, and suggested that the amount of technical effort of post-editing MT output is larger than that of editing TM segments of 75% match. This holds true for four participants in our study (Table 7.12). O'Brien (2006a) made a similar comparison in terms of the cognitive effort in editing English-French and English-German translation using an eye-tracking system, and found that the cognitive load of post-editing MT output was at the same level as that of editing the 80-90% fuzzy match TM segments. If we compare her finding with our results for temporal and technical effort, it holds true for three participants in terms of both temporal and technical effort. Respecting the similarities and differences in results

from both prior studies and our study, we suggest that there is still work to be done on comparison of temporal, technical, and cognitive editing effort between MT output and TM fuzzy matches.

7.7.2 Comparison of multiple regression analysis results

We also attempted to check how well the multiple regression results from post-editing of MT output apply to editing of TM fuzzy matches. Table 7.13 compares the result from Model II in the previous section with the result from the same multiple regression analysis on TM fuzzy matches. Since Complexity Index Scores are given only to MT output, there is no result for this variable in Model VI.

Model:	II	VI
Sample:	MT:	TM:
	All sentences	All sentences
Primary Independent Variable:		
GTM Score (Range: 0 – 1)	1.944 *** (0.08)	1.599 *** (0.22)
Source Text Characteristics:		
ST Length (Range: 1 – 45)	0.054 *** (0.01)	0.053 (0.03)
ST Length^2 (Range: 1-45)	-0.001 *** (0.00)	-0.001 (0.00)
Complex/Compound Sentence (0 or 1)	-0.143 *** (0.04)	-0.071 (0.07)
Incomplete Sentence (0 or 1)	-0.217 *** (0.05)	-0.020 (0.10)
List Item (0 or 1)	-0.176 ** (0.05)	-0.174 (0.10)
Procedure Step (0 or 1)	0.254 *** (0.04)	0.164 (0.09)
UI: non-translated (Range: 0 – 3)	-0.088 ** (0.04)	-0.034 (0.10)
UI: translated (Range: 0 – 2)	-0.210 *** (0.04)	0.486 (0.47)
Complexity Index Score (Range: 1 - 14)	-0.030 ** (0.01)	n/a
PE Operation:		
Supplementation (Range: 0 - 8)	-0.068 *** (0.02)	-0.045 (0.02)
General Term (Range: 0 - 21)	-0.018 ** (0.01)	-0.062 ** (0.02)
Dependency	-0.018 **	-0.062

(Range: 0 - 17)	(0.01)	(0.05)
Rewrite (Range: 0 - 40)	-0.017 *** (0.00)	-0.076 *** (0.02)
Punctuation (Range: 0 - 3)	-0.136 *** (0.03)	-0.161 (0.19)
Number of Visits (Range: 1 - 5)	-0.264 *** (0.03)	-0.516 *** (0.06)
Post-Editor Variance:		
Post-Editor A (0 or 1)	-0.592 *** (0.05)	-0.980 *** (0.11)
Post-Editor B (0 or 1)	-0.817 *** (0.05)	-1.054 *** (0.11)
Post-Editor D (0 or 1)	-0.318 *** (0.05)	-0.390 *** (0.11)
Post-Editor E (0 or 1)	-0.246 *** (0.06)	-0.153 *** (0.12)
Post-Editor F (0 or 1)	-0.134 ** (0.06)	-0.099 (0.12)
Post-Editor G (0 or 1)	-0.534 *** (0.05)	-0.491 *** (0.11)
R-squared	0.48	0.46
Number of cases	2,391	513
Dependent variable (for TM): Logarithm of editing speed (Range: -1.110 to 5.844, SD: 0.809) Unstandardised coefficients are shown in bold face, with standard errors in the parenthesis underneath. Asterisks indicate statistical significance. **: p < 0.01, ***: p < 0.001		

Table 7.13 Regression analysis results of the effect on editing speed (MT–TM comparison)

Many of the independent variables, especially those in the Source Text Characteristics category, lose statistical significance, indicating the factors that affect the amount of editing effort may differ greatly between MT output and TM fuzzy matches. An interesting trend can be observed if we look at the substantive significance of independent variables that hold statistical significance for TM matches; while the coefficients for GTM are lower for TM fuzzy matches than MT output, the coefficients for other variables, namely, General Term, Rewrite, and Number of Visits, are remarkably higher for TM fuzzy matches than MT output. This suggests that the correlation between the amount of editing and the editing speed is weaker for editing of TM matches than for post-editing of MT output, and other factors have a larger influence instead.

7.7.3 Edit frequency

We now compare the actual ingredients of the editing tasks between post-editing of MT output and editing of TM fuzzy matches by examining what parts of speech (PoS) tend to be subject to editing in each case. In an effort to understand the edit frequency in relation to PoS, we distinguished the two types of frequencies: *absolute frequency* and *relative frequency*. Absolute frequency means the percentage of edits on that particular PoS among all edits. Relative frequency means the percentage of edits on that particular PoS among all morphs belonging to that PoS. Absolute frequency is of interest in terms of finding out what PoS are causing most time in the PE task as a whole, since items with high absolute frequencies should occupy a large part of the PE effort spent. Relative frequency is of interest from the linguistic point of view as it shows which PoS tend to be subject to editing, which in turn indicates which PoS are most problematic in either MT translation or in use of TM fuzzy matches.

Table 7.14 and Table 7.15 list the figures averaged over all nine post-editors for post-editing of MT output and editing of TM fuzzy matches, respectively. The first column shows whether each PoS is considered to belong to a content morph or a function morph, as explained in section 4.4.3. The next two columns show the number of each PoS contained in MT output or TM fuzzy matches (common to all post-editors) and the ones in the edited output (averaged over post-editors). The next five columns show what types of change have been carried out on what PoS. ‘Del’ stands for deletion, ‘Ins’ insertion, ‘Sub’ substitution. The reason for having two Subs is to show the number of PoS before and after the edit separately. ‘Total Edits’ column shows the sum of the edits, which counts both of the Subs. The ‘Abs. Freq.’ column shows the absolute frequency, calculated by dividing the number of edits of each PoS by the total number of all edits. The ‘Rel. Freq.’ column shows the relative frequency, calculated by the number of edits of each PoS divided by the sum of the number of morphs belonging to that PoS in MT and PE output. The explanation of each PoS can be found in Appendix B.

Content/ Function	Part of speech	Total Morphs		Edited Morphs					Total Edits		
		MT	PE	Del	Ins	Sub (before editing)	Sub (after editing)	Shift			
Content	Adj	36	28	3	1	13	7	4	28	1%	44%
Function	Adj-dep	2	2	0	0	1	1	0	2	0%	52%
Content	Adv	23	18	2	2	8	3	1	16	1%	40%
Function	Adv-stm	15	10	2	2	6	2	1	12	0%	47%
Function	Aux-verb	411	392	25	17	85	75	15	216	7%	27%
Function	Conj	60	63	5	4	14	19	4	47	1%	38%
Content	Noun	1072	1090	29	27	103	124	113	397	13%	18%
Content	Noun-adjv	77	86	5	9	14	20	4	52	2%	32%
Content	Noun-adv	67	83	3	8	10	21	6	48	2%	32%
Function	Noun-dep	160	109	37	14	54	25	7	136	4%	51%
Content	Noun-Sa	749	738	29	31	95	82	72	308	10%	21%
Function	Noun-suff	100	109	9	10	22	30	6	76	2%	36%
Content	N-proper	223	247	2	8	4	23	13	49	2%	10%
Content	Number	32	34	1	2	3	4	2	11	0%	17%
Function	P-adv	197	194	15	21	49	42	15	143	5%	37%
Function	Adjv-con	15	15	1	1	4	5	1	12	0%	38%
Function	Paren	67	74	0	3	1	5	3	12	0%	9%
Function	P-case	1254	1178	116	83	219	179	118	715	23%	29%
Function	P-conj	72	86	9	22	28	30	3	91	3%	58%
Function	Pref	18	23	1	3	1	5	1	10	0%	25%
Function	Pren-adj	41	45	3	4	5	8	1	21	1%	24%
Function	Pron	73	38	15	4	27	3	5	53	2%	48%
Function	Punc	419	495	7	42	15	56	25	145	5%	16%
Function	Sa-Verb	367	356	35	24	56	54	28	197	6%	27%
Function	Space	0	1	0	0	0	0	0	1		
Function	Symb	58	97	0	18	0	23	1	42	1%	27%
Content	Verb	278	257	22	22	98	77	11	229	7%	43%
Function	Verb-dep	28	32	3	4	11	15	1	33	1%	55%
Function	Verb-suff	56	66	5	7	13	21	4	50	2%	41%
Total		5967	5968	385	393	957	957	463	3155		
Content	43%	2556	2582	96	110	347	361	225	778	36%	22%
Function	57%	3410	3386	289	283	610	595	238	1420	64%	30%

Noun=Noun, Noun-Sa=Stem for Sa-Verb , N-proper=Proper noun, Pron=Pronoun, Noun-adv=Adverbial noun, Noun-adjv=Stem for adjective verb, Number=Number, Noun-dep=Dependent noun, Noun-suff=Suffixal noun, Pref=Prefix, Verb=Verb, Verb-dep=Dependent verb, Verb-suff=Suffixal verb, Sa-Verb=S-consonant irregular conjugation verb , Adj=Adjective, Adj-dep=Dependent adjective, Adv=Adverb, Adv-p=Adverb connected to particles, Pren-adj=Prenoun adjectival, Conj=Conjunction, P-case=Case particle, P-conj=Conjunction particle, P-adv=Adverbial particle, Adjv-con=Adjective verb conjugation, Aux-verb=Auxiliary verb, Symb=Symbols, Punc=Punctuation, Paren=Parenthesis

Table 7.14 Average edit frequencies by part of speech and technical effort (MT)

		Total Morphs		Edited Morphs							
		TM	PE	Del	Ins	Sub (before editing)	Sub (after editing)	Shift			
Content/ Function	Part of speech								Total Edits	Abs. Freq.	Rel. Freq.
Content	Adj	3	3	0	0	0	0	0	0	0%	0%
Function	Adj-dep	0	0	0	0	0	0	0	0	0%	0%
Content	Adv	3	3	0	0	0	0	0	0	0%	4%
Function	Adv-stm	5	5	0	0	0	0	0	0	0%	2%
Function	Aux-verb	73	75	2	0	3	3	0	8	3%	5%
Function	Conj	4	5	0	0	0	0	0	0	0%	0%
Content	Noun	215	216	10	0	37	37	2	87	38%	20%
Content	Noun-adjv	10	11	0	0	0	0	0	1	0%	5%
Content	Noun-adv	19	19	0	0	0	0	0	0	0%	1%
Function	Noun-dep	10	11	0	0	1	1	0	2	1%	8%
Content	Noun-Sa	166	165	7	0	6	6	2	22	10%	7%
Function	Noun-suff	41	39	1	0	2	2	0	6	2%	7%
Content	N-proper	56	42	16	0	2	2	3	22	10%	23%
Content	Number	7	7	0	0	0	0	0	0	0%	0%
Function	P-adv	35	38	1	0	1	1	0	3	2%	5%
Function	Adjv-con	7	6	0	0	0	0	0	1	1%	9%
Function	Paren	39	33	4	0	3	3	1	11	5%	15%
Function	P-case	245	248	5	0	8	8	3	25	11%	5%
Function	P-conj	17	18	0	0	1	1	0	3	1%	9%
Function	Pref	6	6	0	0	0	0	0	1	0%	7%
Function	Pren-adj	8	9	0	0	0	0	0	0	0%	1%
Function	Pron	3	2	0	0	1	1	0	2	1%	43%
Function	Punc	100	99	2	0	1	1	1	6	3%	3%
Function	Sa-Verb	68	73	1	0	4	4	1	10	5%	7%
Function	Space	0	0	0	0	0	0	0	0	0%	0%
Function	Symb	13	16	0	0	0	0	0	0	0%	1%
Content	Verb	50	48	1	0	6	6	1	12	5%	13%
Function	Verb-dep	2	4	0	0	0	0	0	0	0%	0%
Function	Verb-suff	24	24	1	0	2	2	0	5	2%	10%
Total		1229	1225	52	0	80	80	16	228		
Content	43%	529	515	34	0	51	51	8	145	64%	14%
Function	57%	700	710	18	0	29	29	8	83	36%	6%

Table 7.15 Average edit frequencies by part of speech and technical effort (TM)

The first major difference that can be noticed is the ratio of absolute edit frequency between content and function morphs. Both MT output and TM fuzzy matches contain the same ratio of content and function morphs, 43% and 57% respectively, while the ratio of edited content and function morphs reverses between MT and TM; nearly two thirds of the total edits (64%) for MT are on function morphs, which supports the

findings of Groves & Schmidtke on German and French PE⁶² (2009), while in the case of TM, nearly two thirds of the total edits (64%) are on content morphs. A similar ratio applies to the relative edit frequency.

This can be investigated in more detail by looking at the percentage of each edited PoS. The PoS with the highest absolute edit frequency for MT is P-case⁶³, which is a function PoS, whereas in the case of TM, it is Noun, which is a content PoS. The next two highest absolute edit frequencies for MT are found on Noun and Noun-Sa⁶⁴, excluding Adj-dep as there are only two morphs in this category in the entire text and thus may not be suitable for quantitative analysis. Noun-Sa also occupies the same percentage for TM, but also N-proper⁶⁵ has the same percentage, suggesting that product- or company-specific terms and tags are subject to editing more often in TM fuzzy matches than in MT output. The absolute edit frequency of P-case for TM is less than a half of that for MT.

A similar trend can be observed for the relative edit frequency. The majority of the PoS with high relative edit frequencies for MT are function PoS, for example, P-conj⁶⁶, Verb-dep⁶⁷, Noun-dep⁶⁸, and Pron⁶⁹. On the other hand, the majority of the PoS with rather high relative edit frequencies for TM are content PoS, such as, N-Prop, Noun, and Verb. (Although the PoS with the highest relative frequency is Pronoun, the number of the cases may be too few for quantitative analysis.)

Now we turn our attention to the edit frequencies in terms of PE operation types discussed in section 7.5.1. Table 7.16 compares the frequencies averaged over post-editors between MT output and TM fuzzy matches, along with standard deviation (SD).

⁶² They have reported that among function words, punctuation edits were the most frequent for English – German PE, whereas the absolute frequency of punctuation in the present study is very low (5%).

⁶³ Case particle, as described in section 4.4.3.

⁶⁴ The noun-form head word of Sa-verb. The explanation can be found in Appendix B.

⁶⁵ Proper Noun.

⁶⁶ Conjunction particle, as described in section 4.4.3.

⁶⁷ Dependent verb. The explanation can be found in Appendix B.

⁶⁸ Dependent noun. The explanation can be found in Appendix B.

⁶⁹ Pronoun

Variable	MT		TM	
	Mean	SD	Mean	SD
Supplementation	3.6%	0.01	15.8%	0.05
Omission	2.4%	0.01	9.2%	0.05
User Interface	0.7%	0.01	8.2%	0.03
Technical term	3.9%	0.02	23.8%	0.04
General term	18.0%	0.02	19.3%	0.06
<i>Bunsetsu</i> Style	6.5%	0.03	3.6%	0.03
Dependency	13.0%	0.04	3.0%	0.02
Rewrite	36.0%	0.09	3.7%	0.05
Structure Style	13.7%	0.05	12.8%	0.07
Punctuation	2.2%	0.01	0.8%	0.01
Total	100.0%		100.0%	

Table 7.16 Edit frequencies by operation types

The reason for having more edits in Supplementation and Omission categories for TM than MT is because of the nature of TM fuzzy match editing; the post-editors need to alter the translation according to the difference between the original ST and the updated one, for example, a supplementation is necessary in the following case.

Original text stored in the TM:

If the marks do not precisely meet your needs, you can create new ones.

Updated text in the translation file:

If the predefined review marks do not precisely meet your needs, you can create new ones.

Similarly, UI and Technical term edits occur more frequently for TM. For example, a technical term edit is necessary in the following case:

Original text stored in the TM:

The messages of certain employees such as senior managers can be kept separate and reviewed by specially assigned reviewers.

Updated text in the translation file:

The messages of certain employees, called "exception employees", can be kept separate and reviewed by specially assigned reviewers.

Although the low frequency of UI term edits for MT can be explained by the fact that the post-editors were instructed to not change them, the frequency of Technical term edits is also remarkably low. This contradicts one of Krings's finding that in his research project, lexical errors accounted for almost one third of all the errors (Krings 2001). The main reason for this may be because the test set used in this project had been machine translated using well populated Symantec user dictionaries.

On the other hand, structural edits, namely, dependency edits and rewrites, are remarkably less frequent in TM editing. This, together with the observation from the edit frequency based on PoS, suggests that the editing of TM fuzzy matches, especially ones with rather high match ratio (above 75% in the present study⁷⁰), is more about local and lexical changes, or 'parts of the sentence', while post-editing of MT output involves more structural and grammatical changes, or 'construction of the sentence'. This may explain one of the major differences between post-editing of MT output and editing of TM fuzzy matches. From the observation above, it can be said that the editing speed of TM fuzzy matches of match ratios 75% and above is not significantly different when compared to the PE speed of MT, but the *nature* of the changes does differ.

7.8 Summary of Quantitative Analysis

We examined some of the descriptive statistics and then ran various statistical tests in this section. The observations based on each individual independent variable sometimes showed a different effect when combined with other independent variables.

In summary:

- Within- and between-post-editor variance is much higher for PE speed compared to the amount of editing.
- The amount of editing performed during PE moderately correlates with PE speed, and this is common to all post-editors.

⁷⁰ A match ratio can be anything between 1% to 99%.

- This correlation, however, is stronger on simple and complex sentences and weaker on incomplete sentences.
- In general, simple sentences are faster to post-edit than complex sentences, and complex sentences are faster to post-edit than incomplete sentences.
- Document component part *Procedure steps* appears to be faster to post-edit compared to other component parts of the document.
- UI terms slow down PE, not because of the necessity to be edited, but simply because of their presence, and this is more so when the term has been machine translated. This suggests that the post-editors may be distracted by the existence of UI terms, but leaving them in English may help to reduce the distraction.
- The complexity index provided by Systran can indicate the effort intensiveness of post-editing to some extent.
- Some of the PE operation types, namely, Supplementation, Rewrite, and Punctuation edits, proved to have impact on PE speed especially on complex/compound sentences.
- The number of visits affect the within-post-editor PE speed, but not between-post-editor PE speed.
- The amount of technical effort of post-editing MT output can be larger than that of editing TM fuzzy matches above 75%, but the amount of temporal effort is smaller for MT output than TM fuzzy matches above 75%.
- The factors that affect the editing speed differ greatly between post-editing of MT output and editing of TM fuzzy matches.
- Editing of TM fuzzy matches involves more lexical changes compared to post-editing of MT output.
- Post-editing of MT output involves more grammatical changes compared to editing of TM fuzzy matches.

7.9 Concluding Remarks

This chapter presented the findings from the quantitative analyses. It began with describing the analysed data and basic statistical information for the dependent and the primary independent variables. It then moved on to presenting the results to answer our core research questions. RQ1 [How does the amount of editing correlate with the amount of effort in post-editing of English to Japanese MT output?] was answered by showing the correlation between the amount of editing and the PE speed. RQ2 [What characteristics of the English source text have significant influence on the amount of PE effort irrespective of individual traits of post-editors?] and RQ3 [What types of PE operation have significant influence on the amount of PE effort irrespective of individual traits of post-editors?] were answered firstly by showing the relationship between each independent variable and the PE speed, and then by means of multiple regression analysis in order to see the combined effect of all the variables and also the effect of each variable holding other conditions fixed. Some of the source characteristics and PE operation types were found to have an impact on PE speed, but in a rather complex manner. RQ4 [What are the differences between editing of TM match segments and post-editing of MT output?] was addressed from different perspectives in order to have an insight into the difference in editing tasks between MT output and TM fuzzy matches.

Chapter 8 REVISITS, OUTLIERS, AND POST-EDITOR VARIANCE

The purpose of this chapter is to follow up some of the quantitative findings in Chapter 7 and address some ‘how’ questions (Figure 3.1, label [6]). There are three main areas of focus in this chapter: revisits, outlier, and post-editor difference. First, section 8.1 examines what was happening during the revisits in order to investigate what triggered such revisits and how they affected the PE speed. Next, section 8.2 has a closer look at the outlier cases detected by the post-estimation test discussed in Appendix C to find out the causes for exceptionally slow PE speed. Then section 8.3 examines overall differences in GTM scores and PE speed between post-editors by 1) focusing on a special case, that is, non-edited (GTM=1) items, 2) reviewing screen capture video footage, and 3) collating with questionnaire answers. In addition, we will discuss results from the questionnaire so as to have an overview of the facts and opinions of participant post-editors in section 8.4.

8.1 Revisits

We investigated in section 7.5.2 and 7.6 the ‘multiple visits’ issue and observed that: 1) segments visited more often, not surprisingly, were generally slower to be post-edited when compared within each post-editor (*within-post-editor difference in PE speed*), however, 2) some post-editors tended to make more revisits than others, but the tendency of making more visits did not make one a slower post-editor than others (*between-post-editor difference in revisiting behaviour*). In this section, each of these findings will be further investigated by means of detailed analysis of the post-edited text and the timing, purpose, and duration of revisits. As discussed in section 7.7.2, revisits also have shown a significant impact on TM fuzzy match editing, therefore we will include TM fuzzy match segments in the analysis in this section.

8.1.1 Overview of revisits

Table 8.1 shows the descriptive statistics of the number and the types of revisits made by each post-editor. There was a total of 650 records for timing, duration, and textual changes made during the revisits.

We first divided revisit cases into three categories: Modification, No modification, and Comment. *Modification* means that one or more changes have been made to the target text during the revisit, while *No modification* means that the text has not been changed as a result of the revisit, and *Comment* means that the revisit was made solely in order to add/delete/alter one or more comment(s) about the PE task.⁷¹ Among 650 revisit cases, there were 247 Modifications cases, 370 No modification cases, and 33 Comment cases.

Post-editor	A	B	C	D	E	F	G	H	I	Total
Number of revisits	1	0	58	38	209	256	44	15	29	650
Modification	1	0	46	31	72	47	28	10	12	247
No modification	0	0	6	4	129	197	14	5	15	370
Comment	0	0	6	3	8	12	2	0	2	33

Table 8.1 Number and types of revisits by post-editors

8.1.2 General characteristics of PE during revisits

We first focus on the general characteristics of the 247 *Modification* cases. By and large, modifications made during the revisits were simpler and more general than the ones in the first visits. In most cases, a relatively small amount of text was changed involving a single or a few *bunsetsu*. Among 247 Modification cases, there was one case of a tag change, and 12 cases of self-corrections, which were to correct misspellings and typos made by the post-editors themselves during the previous visits. Following the classification of Table 6.2, the 118 of the remaining Modification cases were *bunsetsu*-level changes, most of which involve single term changes. The other 116 were structure-level changes, the majority of which (78 out of 116 cases) were related to

⁷¹ Post-editors were instructed, by the written guidelines and oral instructions, to insert short comments when they were not sure about how to translate or modify the terms or expressions, in order to avoid spending too much time on pondering about the issues. The time taken to insert a comment is included in the PE duration since it is regarded as part of PE effort.

stylistic changes, such as voice conversion, sentence-ending style change, particle deletion, and modal verb alteration, all of which did not involve changes in meaning.

Also, the duration of a revisit was generally shorter than the first visit of the same segment by the same post-editor. We define the duration of a revisit in relation to the first visit as a *relative duration of revisit*, and it is calculated using the following equation:

$$\frac{\text{Duration_of_revisit}}{\text{Duration_of_first_visit}}$$

In over 80% of the cases, the value was less than 1, meaning the duration of the revisits were shorter than the first visits. The median value was 0.52, which suggested that the revisits generally took half the time of the first visit.

8.1.3 Between-post-editor difference in revising behaviour

As can be seen in Table 8.1, Post-editor E and F have made revisits exceptionally more often than other participants. However, as discussed in section 7.5.2, despite the large number of revisits, they are not particularly slow post-editors when all participants are taken into account. This phenomenon may be explained by further investigating the revisiting behaviour and the actions made during the revisits.

The first prominent characteristic of their revisits is, as Table 8.1 shows, that many of their revisits have resulted in making no modifications (61% and 76% respectively). This suggests that they constantly went back to the previously post-edited segments not only because they needed to make changes, but more often for the purpose of general review. The second prominent characteristic is the timing of revisits. We classified the timing of revisits in relation to the first visits into four categories: Less than 1 minute, Later on the same day, and Next day. *Less than 1 minute* means that the next visit was made either immediately or less than 60 seconds after the previous visit, *Later on the same day* means the next visit was made later than 1 minute and within the same day, and *Next day* means literally the visit was made on the next day.

Since there seemed to be no widely agreed upon timing categories for this kind of ‘revisiting’ behaviour, the researcher defined these three categories mainly based on her own experience as a professional translator and a TagEditor user. She divided revisiting roughly into three situations: 1) double-checking or spotting (an) error(s) in the translation during the segment closing operation or soon after that (Less than 1 minute), 2) realising (an) error(s), terminology inconsistency, etc. while editing other segments (Later on the same day), and 3) general reviewing (Next day). This categorisation somewhat corresponds to Shih’s report (2006) on an interview study of translation revision behaviour, in which she found that the most common timing for revision after the draft translation was ‘immediately’, probably in order to check the translation while the memory of ST is still fresh, and the second common time gap was ‘overnight’. Table 8.2 summarises the timing of revisits by post-editors.

	A	B	C	D	E	F	G	H	I	Total
Less than 1 minute	0	0	22	13	20	25	15	5	20	120
Later on the same day	1	0	36	25	18	38	28	10	9	165
Next day	0	0	0	0	171	193	1	0	0	365
Total	1	0	58	38	209	256	44	15	29	650

Table 8.2 Revisits timing by post-editors

It is clearly shown that the post-editors E and F made many of the revisits on the second day while none of the other post-editors did except for one case for the post-editor G.

These two characteristics, namely, performing revisits on the next day, and making no modifications, suggest that the main purpose of the revisits by post-editors E and F might not be to make the changes they already had in mind, but to review the translation as the final process of PE, despite the fact that they were encouraged to complete PE in one pass and avoid multiple visits as much as possible. There are some possible explanations for this. Firstly, they are from the same translation vendor and they had received the same two-week PE training provided by the vendor as part of the vendor’s professional post-editor development plan, thus they might have learned a certain way of post-editing including a ‘final review’ process. Secondly, this might be part of their usual translation process, and they applied the same to PE. In any case, revisits without any changes can cause a significant increase in temporal effort. In fact, the median value of the *relative duration of revisits* (explained in section 8.1.2) that involve no

modifications was 0.31, which is smaller than that involving modifications (0.52), but still can be regarded as a substantial amount of time especially when a large number of such revisits are made.

8.1.4 Types and temporal effort of revisits

Following the PE operation categories developed in Chapter 6, we divided Modification cases into finer level categories, and checked how many cases applied to each category and the relative duration of each category (Table 8.3).

Level	Median value of relative duration	Type	Median value of Relative duration	Number of observation
<i>Bunsetsu</i> -level	0.38	Supplementation	0.25	5
		UI term	0.80	7
		Technical term	0.31	50
		General term	0.44	46
		Stylistic	0.34	10
Structure-level	0.67	Dependency	0.50	12
		Rewrite	1.10	10
		Stylistic	0.66	78
		Punctuation	0.83	16

Table 8.3 Types and temporal effort of revisits

As the table shows, structure level changes, especially rewrite and punctuation edits required more time than other edits in revisits. This somewhat corresponds to the finding in section 7.6, suggesting the similarity between first pass PE and revisits in the amount of temporal effort involved in various PE operation types.

Turning to *bunsetsu*-level changes, we notice that the majority of the edits are related to technical or general terms. For General term edits, all of them, except for three cases, were to alter a word to another with a similar or identical meaning but seemingly more suitable in the context. For Technical and UI term edits, we observed a particular pattern where a post-editor had modified the translation of a word and then changed it back to the previous one; we call this a ‘terminology swing’. Among 57 UI and Technical term edits, 25 were terminology swing cases, and sometimes a term was changed to and fro between two terms more than once in one segment. Sometimes, presumably in order to avoid terminology swings, participants modified the UI or technical term in the first

visit, and visit the same segment again to insert a comment to indicate that they could not confirm the correctness of the translated terms. Many of the terminology swings happened either between a Japanese word and a transliteration (for example, ‘状態’ and ‘ステータス’ for ‘status’) or between a transliteration and an English word (for example, ‘ジャーナリングコネクタ’ and ‘Journaling Connector’). One explanation for this might be that the participant post-editors were not provided with a terminology list; they were advised to assume UI and Technical terms had been properly translated by the user dictionary of Systran. It might also have been the case that they tried to standardise the terms by referencing the other occurrences of the same term within the same document, which had sometimes been translated differently in different places (for example, ‘status’ was translated to either ‘状態’ or ‘ステータス’ by Systran, possibly because both translations are registered to one English word in the dictionary). In any case, editing of these terms can be a sizable burden if left in the post-editors’ remit (Roturier & Lehmann 2009). The accurate translation and thorough standardisation of technical and UI terms may help to reduce the number of revisits.

8.1.5 Section summary

This section focused on the revisit cases, and looked at the between-post-editor differences in revisiting behaviour and the within-post-editor differences in PE speed. As for between-post-editor difference, it was found that the revisiting behaviour differed greatly depending on individual post-editors. For the within-post-editor issues, we observed that the duration of revisits were generally shorter than that of first visits, but still a significant element in entire PE work. In addition, terminology instability seems to be one of the main causes for revisiting, suggesting the importance of terminology standardisation and correct translation of terminology by MT systems.

8.2 Outliers in PE speed

After the quantitative analysis in Chapter 7, we ran several post-estimation tests in order to ensure the validity of the statistical test, which is presented in Appendix C. One of

these was the ‘outlier detection’ test, in which we spotted 105 cases of outliers.⁷² In this section, we will further investigate such cases by first examining the records of time stamps of each post-edited segment to see the types (slow or fast) and the levels of deviation, and next reviewing the screen video footage captured by BB FlashBack to further investigate what happened during PE.

8.2.1 Identifying the outliers to be further investigated

We first looked at the detailed PE history data, and noted that the detected outliers included cases where it was not possible to determine if they were true outliers, thus needed to be excluded from the outlier group. There are two reasons for this: 1) detailed data for some of the revisiting instances have not been recorded by SDL Trados for unknown, probably technical, reasons, thus they might have been mistakenly regarded as exceptionally quickly post-edited segments (33 cases), 2) the recorded time included non-PE operations, such as closing and opening files, due to the misuse of the macro by the participants (5 cases). In addition to these cases, we also decided to exclude all 13 outlier cases for post-editor I since she/he had to turn off BB FlashBack due to a computer performance problem, thus we do not have screen video data to further investigate the outlier cases. Among the remaining 54 cases, 17 are exceptionally quickly post-edited cases, and 37 are exceptionally slowly post-edited cases. We also extracted 20 ‘normal’ cases⁷³ in order to make comparisons between them both quantitatively (8.2.2) and qualitatively (8.2.3). Table 8.4 is the summary of the number of cases by post-editors analysed in this section.

⁷² Outliers, in this case, are the cases where actual PE speed was exceptionally fast or slow compared to the speed predicted by the statistical estimation. For more explanation on outliers, see Appendix C.

⁷³ 20 cases whose actual values are closest to their predicted values.

	Fast	Normal	Slow
Post-Editor A	0	4	5
Post-Editor B	3	1	10
Post-Editor C	3	1	3
Post-Editor D	1	5	4
Post-Editor E	0	4	6
Post-Editor F	1	1	4
Post-Editor G	8	2	4
Post-Editor H	1	2	1
Total	17	2	37

Table 8.4 Number of normal and outlier cases (segments)

8.2.2 Comparison of profiles between outlier and normal cases

We first compared the quantitative information between Fast, Normal, and Slow cases against all sentences, which is summarised in Table 8.5.

	All	Fast	Normal	Slow
Average ST length	14.56	7.94	16.30	9.72
Average GTM score	0.69	0.72	0.57	0.79
Sentence structure ratio:				
Simple sentence	20%	6%	5%	16%
Complex sentence	54%	18%	60%	32%
Compound sentence	4%	18%	10%	0%
Incomplete sentence	22%	58%	25%	52%

Table 8.5 Profiles of fast, normal, and slow PE cases

The first thing one notices is the similarity in average ST length and average GTM score between All and Normal, which is intuitively understandable, and the similarity between Fast and Slow, which may be somewhat counterintuitive. The average ST length is remarkably shorter for Fast and Slow cases compared to All and Normal cases. The average GTM scores are higher not only for Fast cases but also for Slow cases compared to All and Normal cases. This corresponds to the cases that produce variance in the correlation between GTM scores and the PE speed, as shown by the circle in Figure 8.1 below.

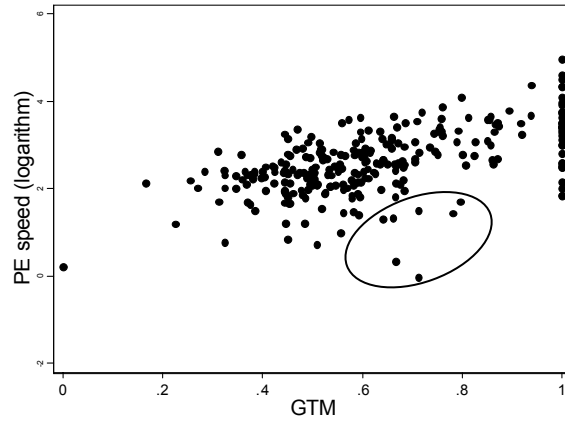


Figure 8.1 Correlation between PE speed and GTM scores (a sample extracted from Figure 7.2)

The fact that the average GTM score is highest for the Slow cases may suggest that the PE was slowed down, for these small number of extremely slow cases, not because of the intensive editing, but for other reasons, which will be further investigated in the next section 8.2.3. In addition, the ratio of incomplete sentences is remarkably higher for Fast and Slow compared to All and Normal. This may partly suggest that the relatively short, incomplete sentences can be either very easy or very difficult to post-edit possibly because sometimes those sentences are simple and clear but sometimes ambiguous due to the lack of context (Aikawa et al. 2007, O'Brien & Roturier 2007).

8.2.3 Qualitative investigation using screen capture video

Next we replayed the screen capture video recorded by BB FlashBack for Fast, Normal, and Slow cases. As mentioned in section 5.1.5, BB FlashBack records computer screen activities including cursor and mouse movements, typing, deletions, copy & paste, choosing menus and options, and so on, and allows one to replay them on other computers. By observing the screen capture video, we can gain some insight into how the PE time was spent on different activities. We roughly divided screen activities into four categories: *editing*, *pausing*, *scrolling*, and *referencing*. **Editing** includes any activities performed in order to edit text, including typing, using arrow keys, moving the mouse pointer, selecting, copying, and pasting text. We define **pausing** as the duration during which no operation is performed. Pause analysis is considered to provide important information about cognitive processing in text production (Schilperoord 1996), and has also been employed in some translation studies (Jakobsen 1998, O'Brien 2006c, Carl et al. 2008, Koehn 2009). Short pauses (less than one to two seconds) are

usually used as delimiters between consecutive text processing actions, while longer pauses are considered to be indicators of cognitive effort. While O'Brien (ibid) introduces the concept 'pause ratio', which is calculated by dividing the total pause duration in editing a segment by the total editing duration of that segment, Koehn (ibid) used the actual time of each pause in seconds to examine the frequency of the pauses of different length. In the present study, however, we are aiming only at having a general picture of pausing behaviour in the outlier cases, instead of performing a detailed quantitative analysis, thus we decided to use simplified categories for pause durations: when it lasted for more than 20 seconds, it is regarded as a *long pause*, and if it lasted for more than one minute, a *very long pause*. The threshold of 20 seconds was determined based on the results from the researcher's own 'time trial.' We picked 20 sentences of lengths between 14 and 25 ST words from the data set, and timed the duration the researcher needed to read and understand the ST and the MT output well enough to start editing. The reasons for using the range 14 to 25 words are that the average ST length of the entire data set is 14.55 words (median is 14), and that 90% of the sentences in the data set are under 25 words. We assume that, if a pause is longer than the time that is required for reading and understanding typical ST-TT sentence pairs, that is an indicator of encountering difficult problem(s) to solve. The maximum time the researcher needed to start editing was 20 seconds, except for the two cases where it took her more than 40 seconds (one case with an ambiguous source sentence, and another with unintelligible MT output). This is a rather arbitrary number, but somewhat coincides with Jakobsen's tentatively identified "cognitive rhythm" of 18 seconds in translation processing (Jakobsen 1998: p.87). The one minute threshold was defined also arbitrarily as another indicator of a higher level of difficulties, which is also used by Koehn for an indication that "the translator is stuck" (ibid: p.255). **Scrolling** means when the screen is moved up or down, while **referencing** means moving outside of Trados TagEditor in order to use some other resources for additional information. There are some other activities that cannot be classified in any of the above mentioned categories. One of the examples is to move a mouse pointer that does not result in any practical action, which has been observed only among a few participants. It seems to be a habitual action of some individuals when focusing on certain parts of the text or pondering, but this cannot be confirmed.

8.2.3.1 Characteristics of Fast cases

Among 17 Fast cases, 11 were GTM=1, meaning no change was made during PE, and four were GTM=0 items, all of which were simply changed back to English (all four were technical terms that consisted of a few words). In both cases, the decisions were made quickly with no pausing, scrolling, or referencing. For the remaining two cases, only a few small changes were made.

8.2.3.2 Characteristics of Normal cases

Unlike the Fast cases, the GTM scores in this group varied greatly. However, there are distinct characteristics of the PE process in this group; during the entire PE time, editing was constantly happening. There were only two cases where a ‘long pause’ was observed either at the beginning or in the end. We interpret these instances as reading the source and/or target sentence to understand the content in the former case, and confirming the changes in the latter case. The pauses occurring at the beginning and at the end of the sentences and paragraphs have also been observed and interpreted similarly by a number of researchers (Jakobsen 1998, Carl et al. 2008, Koehn 2009). There was no scrolling or referencing observed in any of the normal cases.

8.2.3.3 Characteristics of Slow cases

Compared to the Normal cases, the editing in this group occurred sporadically with a number of long and very long pauses, in many cases at the beginning, but also in the middle of the PE process, often following the scrolling up and down, possibly checking the context and/or standard expression in the surrounding segments. Referencing was being performed to search general or technical terms by different means including Web-based dictionaries, the most popular of which being ‘Eijiro’,⁷⁴ Trados Translator’s Workbench Concordance,⁷⁵ and a global search in the source or post-editing text in a text editor. This coincides with the finding of Karamanis et al. that the main cause of the delay in translation process is research using various resources including Trados

⁷⁴ Eijiro is an English – Japanese electronic dictionary project initiated by a Japanese translator in the 1980s, and has been constantly growing since then by a collective effort of the participants. It not only provides definitions of words, but also numerous example parallel phrases and sentences extracted from real-life translation in various domains, though the accuracy is not guaranteed by the authors/editors.

⁷⁵ Users can invoke Concordance tool from Translator’s Workbench to look up word/phrase in translation memory database.

Concordance and online information (Karamanis et al. 2010). One or more long pause(s) occurred in 21 cases out of 37, very long pause(s) in 7 cases, scrolling in 12 cases, and referencing in 8 cases, with some overlaps. Only three cases were free from long/very long pauses, scrolling, or referencing; one with extremely intensive editing, another with a number of cases of mistyping, retyping, and self correction, and a third with the repeated failed attempts to close the segment due to the faulty behaviour of the custom macro.

8.2.4 Section summary

This section investigated the outlier cases by observing the screen capture video footage. There were some distinct characteristics in editing behaviour between Fast, Normal, and Slow PE cases. While editing was constantly happening during PE for Normal cases, it was sporadic in Slow cases. The slowdown of PE was mainly caused not by intensive editing, but by other activities, such as pausing, scrolling, and referencing. This finding somewhat corresponds to what Plitt & Masselot reported from their experiment, in which keyboard activities were recorded, that only 10% of the entire PE time was spent on typing, and the keyboard was not active for the remaining 90% of the time, which was presumably spent on “reading, thinking, and consulting of references” (Plitt & Masselot 2010: p.14).

8.3 General Post-Editor difference

So far, we have mainly focused on the causes for increasing the amount of PE effort independent of individual post-editors. In this section, we turn our attention to the post-editor difference. Figure 8.2 shows the average PE speed and the average GTM scores by post-editor presented in a descending order respectively. PE speed was calculated based on the sum of duration spent on each segment, that is, if a segment had been visited more than once, the PE speed was calculated by dividing the word count of the segment by the aggregated time (minutes) from all visits made to the segment by a given post-editor. The GTM score was calculated based on the MT output and the final PE product.

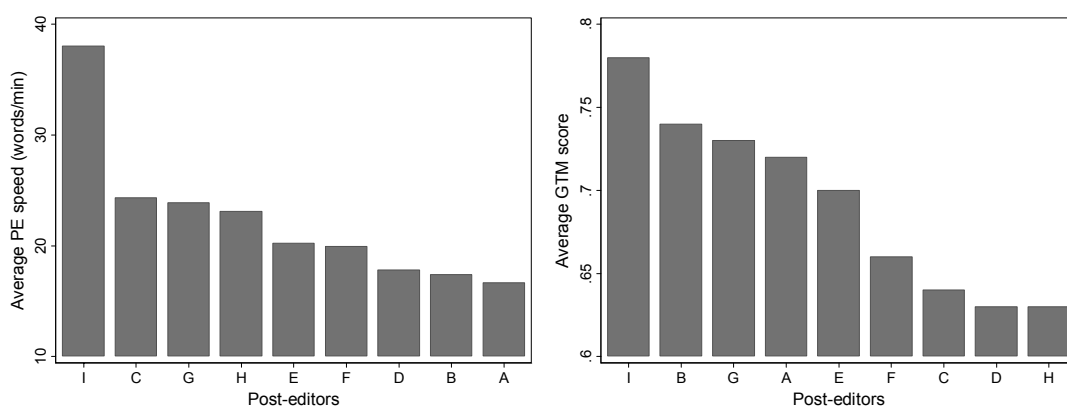


Figure 8.2 Average PE speed and GTM scores by post-editors

The most striking observation from these two graphs is the particularity of post-editor I. She/he is more than twice as fast as the slowest participant, post-editor A, and more than 1.5 times faster than even the fastest of the rest, post-editor C. Yet she/he is the one who has made the largest amount of editing in total. We examined the questionnaire answers to discover if there was any outstanding characteristics about this post-editor. First of all, post-editor I has the longest and broadest experience in translation. As presented in Appendix D, we asked the participants about their translation experience in three fields: computer software documentation, technical documents other than computer software documentation, and anything other than computer software or technical documentation. The seven out of nine participants had more than five years of experience in one or two of these fields, but post-editor I was the only one who had more than five years of translation experience in all three categories. The rich and broad experience in translation may affect not only the PE speed but also the amount of editing; experienced translators tend to make more changes to raise the final quality of the target text (de Almeida & O'Brien 2010). She/he was also one of the three participants who had more than five years of experience in using TagEditor, which may have helped with throughput. Post-editor I might have been an interesting case to investigate the actual PE practice by means of screen video footage, though as mentioned previously, she/he had to turn off BB FlashBack due to a computer performance problem, thus unfortunately we cannot investigate her/his practice using the screen video footage.

For the rest of the participants, both PE speed and GTM scores have certain ranges of values, though the average PE speed and GTM scores do not seem to have direct

relationship when looked at from a post-editor by post-editor basis. In the following subsections, we will investigate if such differences can be explained by reviewing the non-edited items in section 8.3.1, screen capture video in section 8.3.2, and questionnaire results in section 8.3.3.

8.3.1 Non-edited segments

Non-edited segments are identified by a GTM score of 1, which means no textual change has been made between the MT output and the final PE product. The agreement level among post-editors on which segments to leave intact is noticeably low; only six segments were left unedited by all nine post-editors, 10 additional segments were left unedited by eight, another nine segments by seven post-editors. Table 8.6 shows the number of unedited or GTM=1 cases and the average PE speed for those segments by post-editors.

Post-Editor	Number of GTM=1 cases (SD: 17.84)	Average Speed (SD: 13.23)
Post-Editor A	69	28.42*
Post-Editor B	86**	23.30**
Post-Editor C	47*	54.88*
Post-Editor D	42*	37.98
Post-Editor E	60	34.28
Post-Editor F	45*	40.42
Post-Editor G	61	45.97
Post-Editor H	49	45.67
Post-Editor I	91**	66.35**
Asterisks indicate that the figure is significantly different from the average (based on t-tests). *: p<0.05, **: p<0.01		

Table 8.6 Number of GTM=1 items and average PE speed

The number of GTM=1 cases differs greatly depending on post-editors and so does the average post-editing, or more precisely in this case, ‘reading and verifying’ speed for these segments. Figure 8.3 compares the average PE speed for all segments and the average PE speed for GTM=1 cases by post-editors.

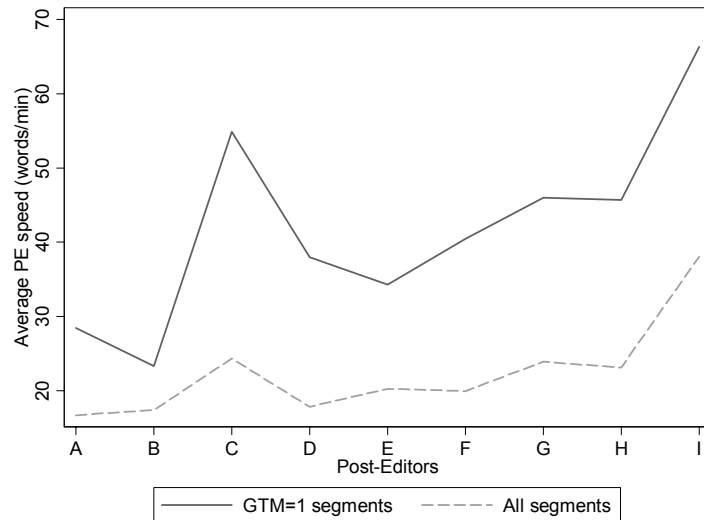


Figure 8.3 Average PE speed for GTM=1 segments and all segments

In general, fast post-editors in terms of overall average PE speed are also fast in reading and deciding that a given segment needs not be edited. However, as the lines show, the between-post-editor difference is larger for PE speed for GTM=1 cases compared to overall speed. This may suggest that the speed for reading and deciding on a PE strategy is a stronger factor than the editing speed in differentiating the general PE speed between post-editors.

Now we turn our attention to the relationship between the number of GTM=1 cases and the average PE speed for those segments by post-editors, which is summarised in Figure 8.4.

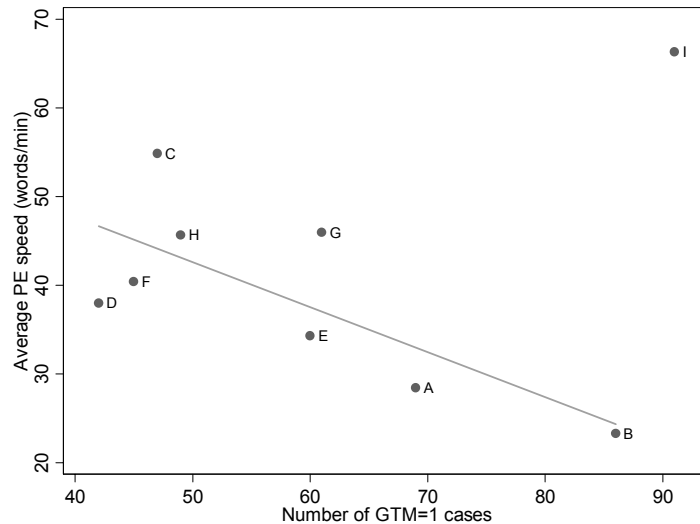


Figure 8.4 Number of GTM=1 cases and average PE speed

The figure indicates that there is a negative relationship between the number of unedited segments and the average PE speed for them, except for post-editor I, who is an outlier in a number of aspects throughout our analysis.⁷⁶ The overall relationship indicates that the post-editors who leave a larger number of segments unedited spend more time on deciding to do so than the post-editors who leave a smaller number of segments unedited. In other words, those who make changes more often may do so partly because they tend not to spend a long time on contemplating whether they should leave the segment unedited compared to the ones who (make an effort to) leave more segments intact. This, combined with the result from the comparison between overall PE speed and PE speed for GTM=1 cases, may suggest that the slower post-editors are so partly because they are being more cautious about making changes than the faster ones.

8.3.2 Observation of BB Flashback

In observing the screen capture video, we focused mainly on the three points: 1) screen layout, 2) use of mouse and keys, and 3) use of information resources, and checked if there is any relationship with these aspects and the average PE speed or GTM scores. Unfortunately, as previously mentioned, we do not have screen capture data for post-editor I, thus the discussion is made based only on the remaining eight participants.

⁷⁶ The Pearson correlation coefficient is -0.74 ($p=0.03$) when excluding post-editor I. The fitted line also shows the relationship excluding post-editor I.

8.3.2.1 Screen layout

All participants, except post-editor A, arranged SDL Trados TagEditor and Translator's Workbench so that both are fully visible, with or without some other applications, such as Windows Explorer and online dictionary. Figure 8.5 shows an example of such a case. Post-editor A, who placed Translator's Workbench behind TagEditor, brought up Translator's Workbench when necessary. By examining the questionnaire answers, it was found that this participant was the only one who had never used SDL Trados TagEditor before. Post-editor A is also the slowest participant, but since there is only one case, we cannot generalise the relationship between the experience with the tool, screen layout, and the average PE speed.

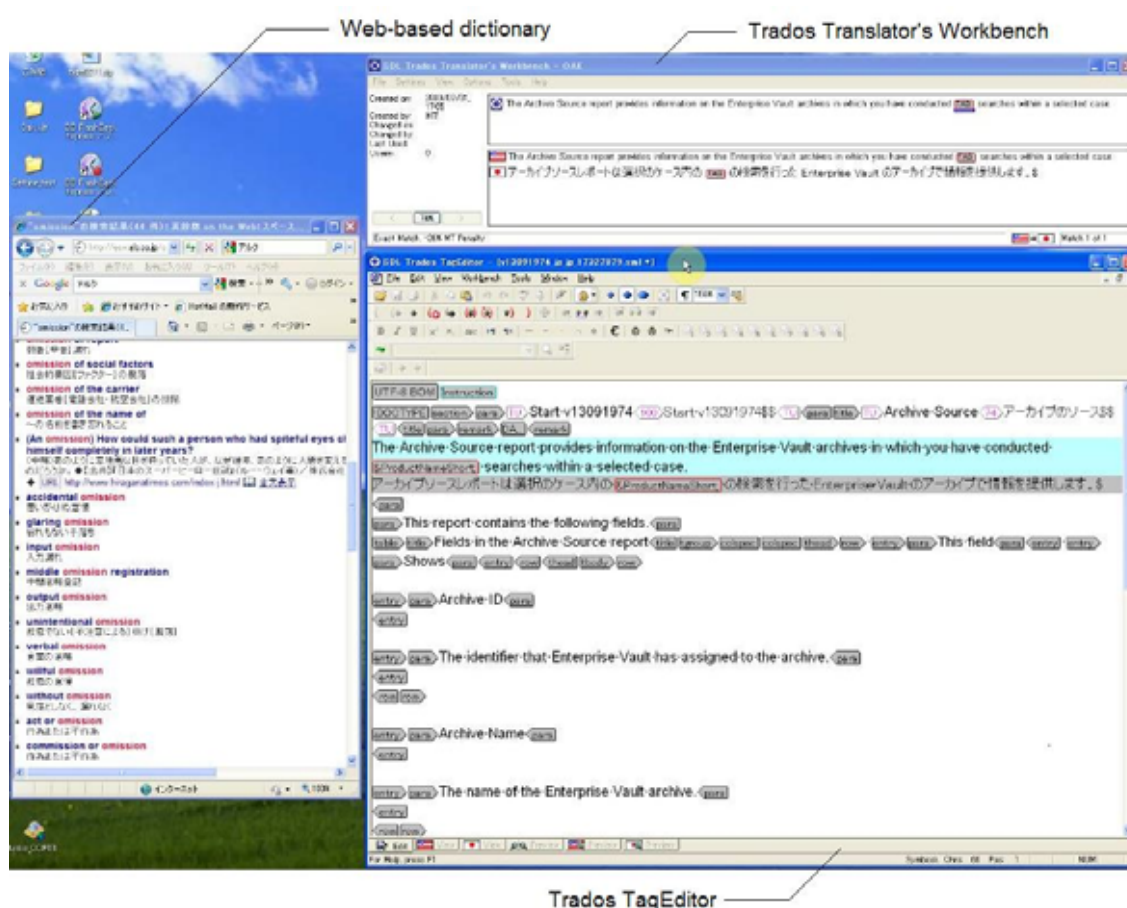


Figure 8.5 Example of screen layout

Translator's Workbench is informative when editing TM fuzzy matches as it shows which part of the ST has been modified, along with the fuzzy match ratio, so that the translators/editors can have an idea about how much editing is necessary on what part.

When post-editing MT output, on the other hand, the benefit is less as it does not show extra information such as the possible accuracy of the MT output.⁷⁷ However, the participants were informed that the target text was a mixture of MT output and TM fuzzy matches, and it was distinguishable by looking at the ‘Created by:’ field on the left side of Translator’s Workbench (Figure 8.5). In addition, the answers to a question in the questionnaire regarding the use of TagEditor tells us that eight out of nine participants constantly reference the ‘Created by’ field. These results may indicate that translators find it helpful to obtain information about the origin of the translation when editing it, which supports the finding of Karamanis et al. that the translators consult the information about the origin of translation in gauging the trustworthiness of the translation (Karamanis et al. 2010).

8.3.2.2 Use of mouse and keys

The ratio of the mouse/key use while editing varied greatly. While post-editors C, D, G, and H had their mouse pointer mostly or often fixed somewhere on the screen (meaning the mouse is being unused) and the majority of the insertion point movements and text selections were made using the combination of Shift, Ctrl, and arrow keys, post-editors E and F had their mouse pointers fixed for a shorter time and used the mouse slightly more often for insertion point movements and text selections, and post-editors A and B almost never had the mouse pointer fixed; the mouse was constantly used even for short distance insertion pointer movements and the word-level text selections. All four participants who tended to avoid using the mouse (C, D, G, and H) had three or more years of experience in using TagEditor, and three of them are among the fastest post-editors (Figure 8.2), which may suggest some relationship between the experience in TagEditor, use of key/mouse combination, and the PE speed, although the direction of the causation is not clear.⁷⁸ In terms of the relationship between key/mouse usage and GTM scores, however, no correlation can be seen.

⁷⁷ A related technology, ‘confidence index’, has been suggested by several researchers, as mentioned in section 2.4.1.

⁷⁸ This partly corresponds to the result of the study by de Almeida & O’Brien (2010), in which they reported that the two fastest participant post-editors were the ones who used keyboard more often than any other participants.

8.3.2.3 Use of information resources

Three types of information resources were utilised among eight participants: 1) Web-based dictionaries by five participants: A, B, D, E, and F, 2) SDL Trados Concordance by three participants: C, E, and F, and 3) source and target text file with a global search function of a text editor by one participant: C. The frequency of the use of such resources was, however, not very high. Most post-editors used these resources from a few to less than 20 times. Trados Concordance and the global search function seemed to be used for looking up project-specific translations of certain terms, such as ‘automatic categorization’ and ‘other user access’. Similar to the findings of Karamanis et. al. (2010), these are not uncommon words; the participants, as they are experienced translators, most likely knew the meaning of these words, but needed help to find out how these terms had been translated in other parts of the same document. However, the relatively small number of Concordance users (three) contradicts the finding of Karamanis et al. (ibid) that a Concordance search is quicker than using online resources, thus usually the first step in terminology lookup. A possible reason for this is that we made clear to participants that majority of the segments were machine translated, thus probably the participants did not expect to find solutions from Concordance. On the other hand, Web-based dictionaries, most frequently Eijiro, were used by five participants for looking up not only technical terms, such as ‘purge’, ‘custodian’, ‘production run’, and ‘Bates number’, but also general words and phrases, such as, ‘accidentally’, ‘cap’, ‘quick navigation’, and ‘best chance to’. Again, it is difficult to believe that professional translators do not know the meaning of these general words. Based on this assumption, together with the characteristics of the Web-based dictionaries they use (as mentioned in section 8.2.3.3, Eijiro offers a large number of real-life sample translations from various genres with no guarantee for accuracy), we speculate that the dictionary lookup for general words and phrases is *not* to check the meaning of them, but to obtain a list of possible translations for them. As Désilets et al. point out, participants may have cared “more about recall than precision” in word/phrase lookup (Désilets et al. 2008: p.341). The most frequent user of online resources was post-editor B, who looked up over 50 words in Eijiro during the whole PE task, including a number of non-technical, idiomatic phrases, such as ‘in effect’, ‘have no use for’, and ‘find useful’. This participant has 3-5 years’ experience in translation of software manuals and more than five years’ experience in translation of

technical documents other than software manuals, thus again it is unlikely that the frequent dictionary lookup was necessary due to the lack of knowledge in English words. The five participants who made use of online dictionaries, namely, A, B, D, E, and F, are the five slowest post-editors (Figure 8.2), which may suggest some relationship between the frequency of dictionary lookup and the PE speed; dictionary lookup may slow down PE, or frequent dictionary lookup, due to the lack of active vocabulary, (over)carefulness in choosing target words, or some other reason, is one of the characteristics of relatively slow post-editors.

8.3.3 Background data from questionnaire

As discussed in sections 5.5, we included in the questionnaire some questions about the participants' experience in translation, post-editing, and SDL Trados TagEditor. In this section, we examine if there are any relationships between the length of experience in these fields and the PE speed or GTM scores, which are summarised in Figures 8.6, 8.7, and 8.8 below. The average PE speed (left) and the average GTM scores (right) are shown in a descending order respectively. The years of experience are classified into five categories: no experience, less than one year, one to three years, three to five years, and more than five years, and shown by diamond shaped dots.

8.3.3.1 Experience in software manual translation

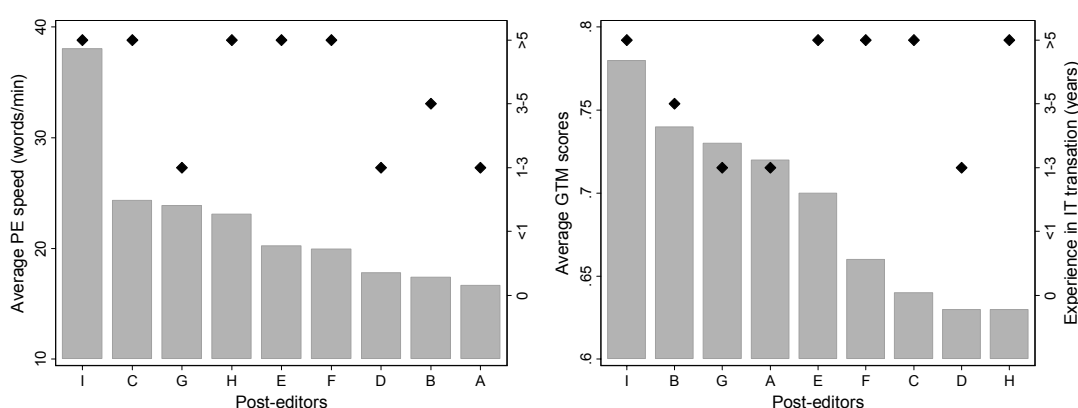


Figure 8.6 Experience in software manual translation and PE speed / GTM scores

We included questions about the participants' experience in translation of software manuals, technical documents other than software manuals, and non-technical domains.

Since all the participants had some experience in software manual translation, we will focus on the experience in software manual translation here. Figure 8.6 seems to imply a loose relationship between the years of experience in software manual translation and the PE speed / GTM scores. The dots at the highest positions, which means more than five years of experience, tend to gather on the left side of the left graph (faster PE speed) and on the right side of the right graph (lower GTM score), which suggests that the participants with more experience in software manual translation tend to make more changes during PE but are faster to PE than others, which coincides with the findings by de Almeida & O'Brien (2010).

In order to check the statistical significance, we divided the nine post-editors into two groups: A) the ones with more than five years of experience, and B) the ones with five years or less experience, and ran a t-test. The mean PE speed for the more experienced group was 18.94 words/min (SD: 1.66), and the less experienced group was 25.13 words/min (SD: 3.33). The mean GTM scores for the more experienced group was 0.71 (SD: 0.03), and the less experienced group was 0.68 (SD: 0.03). The results of the t-test, however, did not prove statistically significant in either relationship with p-values of 0.17 and 0.57 respectively.

8.3.3.2 Experience in software manual post-editing

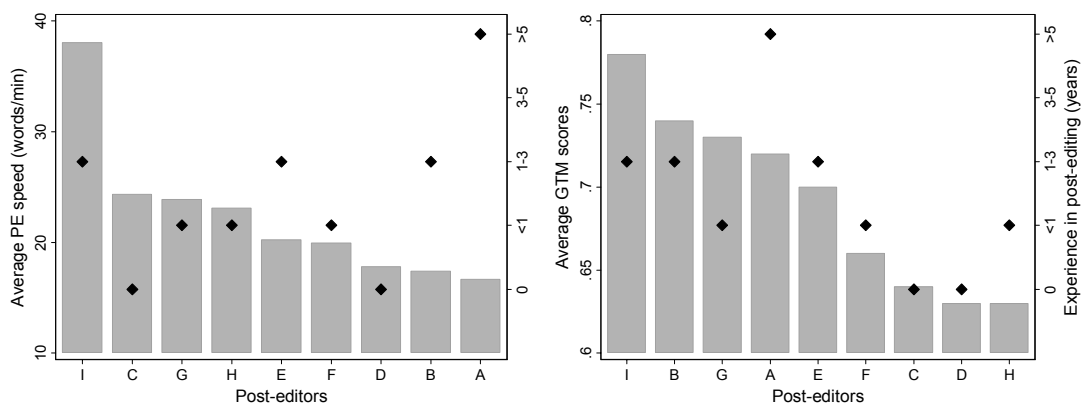


Figure 8.7 Experience in PE and PE speed / GTM scores

As for post-editing, we also asked participants how many years of experience they had in post-editing software manuals, other technical domains, and non-technical domains. Seven out of nine participants had some experience in post-editing software manuals,

and one had less than one year's experience in post-editing non-technical documents. (Although we asked the vendors to provide only participants with some experience in PE, the result of the questionnaire revealed that one participant had no experience in PE.) We focus also on PE experience in software manuals here. The left graph in Figure 8.7 does not seem to show any relationship, but the right graph suggests that the participants with less experience in post-editing software manuals tend to make more changes thus resulting in lower GTM scores than others.

We divided the nine post-editors into two groups again: A) the ones with more than one year of experience, and B) the ones with less than one year's experience, and ran a t-test. The mean GTM scores for the more experienced group was 0.74 (SD: 0.17), and the less experienced group was 0.66 (SD(0.17). The t-test resulted in a p value of 0.02, which is not very strong, but shows some statistical significance.

8.3.3.3 Experience in TagEditor

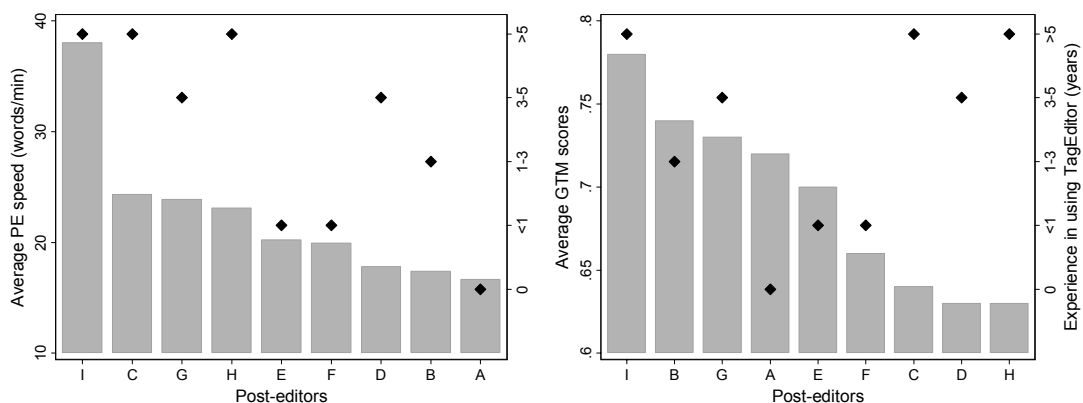


Figure 8.8 Experience in TagEditor and PE speed / GTM scores

Eight out of nine participants had some experience in using TagEditor; all of them had used it as a translation tool, while four of them had used it also as a PE tool. Figure 8.8 shows the years of experience in using TagEditor regardless of the purpose. The figure suggests no particular relationship between the TagEditor experience and the average GTM scores (right graph), but it seems that the participants with more experience are faster in PE in general (left graph). We again divided the participants into two groups: A) the ones with more than three years of experience, and B) the ones with less experience. The mean PE speed for the more experienced group was 25.43 words/min

(SD: 3.36) and the less experienced group was 18.56 words/min, but the result of the t-test was $p=0.12$, which did not prove statistically significant.

8.3.4 Section summary

This section examined the post-editor difference by reviewing non-edited segments, screen capture video, and answers to the questionnaire. We found a relationship between the number of non-edited segments and the PE speed for those segments on a post-editor-by-post-editor basis; the post-editors who leave a larger number of segments unedited spend more time on deciding to do so than others. It was also found that the participants who had relatively long experience in using TagEditor tended to use the keyboard more often than the mouse, and their PE speed was generally higher than others. The use of reference resources also differed between participants, and the ones who made use of online dictionaries were also relatively slow post-editors, though the causation was not clear. The participants with relatively long experience in software manual translation seemed to make more changes but were faster to post-edit than others, though statistical significance was not demonstrated. The participants with no or relatively short experience in software manual post-editing seemed to make more changes, with a rather weak statistical significance. The participants with relatively long experience in TagEditor appeared to be faster in post-editing than others, though again the statistical significance was not proven.

8.4 Overview of questionnaire results

In this section, we will summarise the answers to some other questions in the questionnaire, which asked about their experience in PE training, methods of PE, and opinions about PE training, PE guidelines, and PE in general. However, answers to the question about PE guidelines contained mainly opinions about other problems regarding PE efficiency, thus we will introduce them here as general problems that cause PE inefficiency.

8.4.1 PE training

Out of nine participants, only three had received PE training, all of them given by the translation vendors as internal training sessions. Two of them had taken a two-week

training course, and one a course over four days. What they found most useful about the PE training included that they learned common problems of MT output and how to efficiently correct errors.

We also asked all participants what kind of training they thought would be useful. The contents of the preferred training can be categorised into three types of knowledge: 1) MT system and MT output, 2) desired quality of final products, and 3) post-editing methods. The questionnaire was answered in Japanese, and the following are some of the answers translated into English by the researcher.

About MT system and MT output:

- *Explanation about how MT systems work and how they process translation*
- *Common patterns of MT output, including sentence structures*

About the desired quality of target language:

- *Clear guidelines about the desired level of translation*
- *Desired Japanese standard, more specifically, acceptable level of 'unreadability' by readers and/or document producers, shown with lots of examples of before and after PE*

About post-editing methods:

- *Reviewers who are used to correcting human translation tend to make too many corrections, so we need concrete and detailed suggestions about how we should address PE.*
- *The difference between translation and PE*
- *Domain specific PE training*

8.4.2 PE methods

We asked post-editors to explain in detail how they usually proceed with PE tasks, such as whether they work monolingually or bilingually, or which text they read first, ST or MT output, etc. None of the nine participants worked monolingually; they all read both ST and MT output. However, the order of reading depended on the individual participants. Three out of nine participants read the MT output first, while six read the

ST first. Among those who read the MT output first, two then read the ST, and started editing the MT output matching it against the ST. The remaining one, after reading the MT output, judges whether the MT output is intelligible monolingually. If it is, the ST is read and corrections are made if necessary. If it is not, the ST is read, the new translation is roughly constructed in the mind, and the target sentence is rebuilt. Among those who read the ST first, four then read the MT output and start editing, while two first perform translation in their mind, then read the MT output. Two out of nine participants mentioned that they strive to reuse the MT output as much as possible by reshuffling the existing parts of MT output when producing the final translation. Mossop (2007) observes that when people review translation, some read the translation first and then the source text, while others do otherwise, and he poses a question whether one of the two methods produces better results than the other. In our study, we examined the reading order with the PE speed and the GTM score, but found no relationship between them.

8.4.3 Reason for inefficient PE

We were interested in listening to the participant's opinions about PE guidelines, not particularly about the guidelines used in the present study, but PE guidelines in general, to see if there are any particular aspects of guidelines they find difficult to follow or are causing problems in PE. Six out of nine participants answered that there are problems in guidelines in general, however, when they elaborated, they mainly focused on issues regarding the glossary, TM, MT output, and ST as follows:

- *Nonstandardised terminology in the glossary and TM*
- *Varied format specifications for UI terms, product names, and so on*
- *Nonstandardised use of symbols and spaces in the target text*
- *Nonstandardised sentence ending style of the target text*
- *Source text ambiguity*

8.4.4 MT and PE in general

We asked the participants whether they find post-editing of MT more difficult or problematic than editing of TM fuzzy matches. Two out of nine participants answered TM fuzzy matches are easier to edit than MT output in general, while none said MT

output is easier than TM fuzzy matches, but the other seven participants answered it depends on the quality of the MT output and/or the TM fuzzy match ratio. Since we only used the TM fuzzy match segments above 75% match, this result implies that at least some MT outputs were easier to edit than 75% TM fuzzy matches. This agrees with our quantitative findings discussed in section 7.7.1.

Seven out of nine participants gave some opinions about MT and PE in general and about the participation in the study project. Two of them gave positive comments about the quality of MT output, only one expressed disapproval of using MT systems. Three gave specific opinions about possible improvements and suitable usage of MT systems. Four gave positive comments about the participation in the study project, while also four expressed difficulties encountered during the experiment. Many of the comments included detailed observations and remarks, which may indicate their interests in PE tasks. The following is the translation of some comments.

About MT:

- *I was surprised at the level of MT output. Many of the short and simple sentences did not need any editing.*
- *The quality of the MT output was much higher than I thought, and would streamline the translation production. However, complicated sentences with commas, dashes, and conjunctions were often translated into unintelligible Japanese, which could increase the time and trouble of translation production.*
- *Most of the pronouns and sentence-endings needed editing, and it was sometimes easier to retype than reordering the words and phrases of MT output to produce correct sentences; there is no advantage in using MT systems.*
- *MT may be best utilised as a means of terminology standardisation and block-by-block translation [rather than as a tool for producing complete translation].*
- *User dictionaries could be better populated.*
- *Sometimes two similar source sentences are machine translated into very different target sentences, which compromises the benefit of using MT systems.*
- *I think MT systems are suitable for instruction manuals.*

About PE and the study project:

- *Sometimes I felt I made too many corrections to a target segment, and on a rebound, I tended to make too few corrections to the next segment, leaving an awkward Japanese sentence untouched, resulting in quality instability.*
- *It is psychologically tough to be required to keep producing ‘second quality’ Japanese, which may cause problems in quality and efficiency. This issue needs to be addressed [when deploying MT systems].*
- *Understanding the procedure of the experiment was more time-consuming than actual editing.*
- *I had trouble distinguishing UI terms from other proper nouns since there was no guide [glossary].*
- *The participation in the study project helped me to understand post-editing of MT output.*
- *I think this is an interesting study both from practical and academic points of view.*
- *The project was only for two days, but I felt like doing some more.*

8.4.5 Section summary

This section reviewed answers to some of the questionnaire questions to know more about the participants’ background and opinions. Only three out of nine post-editors had had PE training before, while others appeared to be interested in having training. The desired skill sets for post-editors have been discussed since the 1980s, such as knowledge in the subject domains, language competence, and computer skills (Vasconcellos 1987, Wagner 1987 etc.). However, our participants already had these expertise and skills as professional translators, and were seeking more specific knowledge. They were not only interested in the efficient methods of PE, but also about MT systems themselves, which supports O’Brien’s suggestion on post-editor training (O’Brien 2002). It was also insightful that participants mentioned that they wanted to know the required or acceptable quality levels of PE. This demonstrates two things. Firstly, PE is still in a developing phase, and there is no widely agreed-upon quality level set for post-edited products. Secondly, and more importantly, the post-editors, who are often translators thus are used to producing high-quality target text, are now prepared to adjust the quality of the final text they produce based on various

requirements. As surveys and interviews by TAUS (TAUS 2010) show, the ‘resistance’ from the translators towards PE is often regarded as one of the issues in deploying MT. But the answers to our questionnaire suggest that a flexible and down-to-earth attitude towards PE is the trend. This agrees with the result from Garcia’s PE study on English to Chinese SMT output, in which eleven out of fourteen participants had positive impressions about the PE task; they even thought they would do better when post-editing than translating from scratch (Garcia 2010). Based on these findings, we suggest that (potential) post-editors are becoming more positive about MT + PE workflow, and experienced translators, with their linguistic competence and professionalism, can even play a key role in discussing and establishing quality levels for MT + PE product.

We also asked how they go about performing PE. The methods somewhat varied between post-editors, but none of them worked monolingually. Some of the biggest issues that cause more effort and problems in PE in general, according to participants, include nonstandardised elements, such as terminology, use of symbols, and styles. Some of the participants had positive impressions and constructive opinions about MT output and the PE task, suggesting their interest in this activity.

8.5 Concluding remarks

This chapter served as an explanatory phase to follow-up the quantitative findings in Chapter 7 and to examine the possible post-editor variance. It began by qualitatively investigating the revisit and outlier cases by investigating the time and text data in detail and reviewing the screen capture video, respectively. Some findings were made about individual differences in revisiting behaviour and the characteristics of the editing behaviour during revisits. We then moved on to look at the general post-editor differences by reviewing the non-edited cases, screen capture video, and the answers to the questionnaire. It was indicated that some of the background and editing behaviour of each participant might affect PE speed, though statistical significance was not proven. It also summarised the results of the questionnaire, which revealed some of the participants’ thoughts and opinions towards PE activities.

Chapter 9 CONCLUSIONS

9.1 Aims of the Study

The aim of the present study was to answer the fundamental question ‘What determines the amount of PE effort?’ in order to obtain knowledge to help with streamlining the MT workflow in commercial settings. Answering this question inevitably demands multiple research perspectives as it involves, on the one hand, understanding human behaviour in the editing task, which is often complicated and subject to irregularity, thus difficult to quantify, and on the other hand, obtaining quantifiable results to render the findings informative and useful to the industry. This requirement entailed the necessity to devise a research design that could combine both quantitative and qualitative methods. Various technologies and techniques, such as AEMs, CAT tools, and macros, enabled standardised measurements, while statistical methods, mainly multiple regression, provided a way to quantify complex human activities involved in post-editing. Likewise, different qualitative research methods, including manual classification, questionnaire, and PC screen video observation, were employed to address qualitative aspects of the study. The primary interest in answering the above-stated question was to uncover the common phenomena in PE activities independent of individual post-editors, since understanding post-editor-independent phenomena is more beneficial to MT users. The foci were on measuring the effect of ST characteristics and PE operation types on the PE speed. In addition, we aimed at forming a general idea about the differences in post-editing of MT and editing of TM fuzzy matches, as combining the two technologies has increasingly been the norm in IT localisation today. We also explored and strived to explain some of the significant post-editor variance as individual differences do exist, as with any other human activity, and need to be addressed in order to gain a fuller insight into the PE activities.

9.2 Findings of the Study

The answer for our fundamental question ‘What determines the amount of post-editing effort?’, naturally, spans across different areas. We found that the amount of editing had the highest impact on the PE speed, but also various ST characteristics, including the presence of UI terms and type of sentence structure, and certain PE operation types and actions, especially punctuation edits and revisits, had a high impact on PE speed. It was also found that extreme PE slowdown was caused mainly not by a large amount of editing, but rather by pausing (thinking?) and referencing. In addition, the between-post-editor difference was found to be large in terms of speed, methods, and behaviour.

The following is the summary of findings corresponding to each research question. The main findings of the present study are related to quantifiable aspects of the PE activities (RQ1, 2, 3, and 4), which is supplemented by qualitative findings (RQ5 and Follow-up Issues).

[RQ1: How does the amount of editing correlate with the amount of effort in post-editing of English to Japanese MT output?]

The amount of editing performed during PE was found to moderately correlate with the amount of PE effort (measured by PE speed throughout the present study) on a segment-by-segment basis, and this was observed to be common across all post-editors. However, the correlation was stronger for simple and complex sentences and weaker for incomplete sentences. In addition, it was found that both within and between post-editor variance is much higher for PE speed compared to GTM scores, meaning that the amount of editing performed during PE is similar within and between post-editors, while the time taken to make the changes varies greatly both within and between post-editors. Although stylistic changes can be minimised by instructing post-editors by guidelines and training, corrections in order to make MT output accurate and understandable cannot be reduced by the effort of post-editors; the quality of MT output needs to be improved, which may involve various elements including MT architecture itself and customisation effort, such as user dictionaries.

[RQ2: What characteristics of the English source text have significant influence on the amount of PE effort irrespective of individual traits of post-editors?]

Sentence structure had an influence on the PE speed; simple sentences were the fastest, complex sentences came next, and incomplete sentences were the slowest to post-edit. We also observed some relationships between document component parts and the PE speed; procedural text was faster to post-edit compared to other text types. The ‘sentence complexity’ measured by Systran, which considers various characteristics of a sentence, including sentence length and the number of clauses, conjunctions, and prepositional phrases in a sentence, showed a relationship with the PE speed; the higher the complexity score, the slower the PE speed was. In addition, the presence of UI terms in a sentence was found to slow down PE, and this was not because of the necessity to be edited, but simply because of their existence. Some of the ST characteristics that cause more PE effort could be addressed in the authoring phase by means of CL rules, tagging of certain terms, and so on.

[RQ3: What types of PE operation have significant influence on the amount of PE effort irrespective of individual traits of post-editors?]

The answer to this question was addressed based on our own taxonomy of PE operations. Some PE operation types, namely, supplementation (adding text to express a concept that did not exist in the MT output), rewrite, and punctuation edits, were seen to have a relatively high impact on the PE speed, especially on complex/compound sentences. Among the three, punctuation edits had a particularly high impact indicating the effort intensiveness of editing punctuations despite the fact that the amount of textual change involved is relatively small. Though better quality MT output will help reduce the necessity for those time-consuming PE operations, we suspect that supplementing extra information or inserting/deleting certain types of punctuation may be difficult to handle for RBMT as these edits often require knowledge about the real world and flexible application of writing conventions.

[RQ4: What are the differences between editing of TM match segments and post-editing of MT output?]

The quantitative and qualitative comparisons between editing of TM match segments and post-editing of MT output revealed some important similarities and differences between the two. In terms of effort intensiveness, the editing speed of TM fuzzy matches of match ratios 75% and above was not significantly different when compared to the PE speed of MT. However, the factors that affect the editing speed differed greatly between the two; while post-editing of MT output involved more grammatical changes compared to editing of TM fuzzy matches, editing of TM matches involved more lexical changes compared to post-editing of MT output.

[RQ5: How do different attributes, techniques, and/or behaviour of post-editors affect PE practice?]

The questionnaire was the primary source for answering this question, with the screen capture video as the secondary source of information. We observed some relationships between the participants' varied experience, the PE speed, and the amount of editing made during PE. In general, 1) the participants with less experience in IT-related PE made a relatively larger amount of editing than others, 2) the participants with more experience in TagEditor used the keyboard, rather than the mouse, more extensively, and were generally faster in PE than others, and 3) the participants with more experience in IT translation tended to be faster in PE and made more edits compared to others, though these findings did not demonstrate strong statistical significance.

We learned, from the quantitative analysis phase, that two of the nine participants had made revisits to segments noticeably more frequently than others. The reason for this was further investigated. It was found that many of the revisits made by these two participants resulted in no modifications, which suggested that those revisits were made not because they knew that further changes were necessary, but more likely as the general final review process. Interestingly, the general PE speed of these two participants was not particularly slow compared to others. However, considering that the relative duration of the revisits with no modification was still 30% of that of the first visits, such seemingly strategic or habitual revisits could be considered to slow down the PE process of individual post-editors unnecessarily when they were advised to avoid inessential revisits in order to spend as little time as possible on PE.

[Follow-up Issues]

Along with the effort for answering the actual RQs, we also further examined some of the instances where the amount of PE effort was particularly high, namely, revisits and outlier cases. The quantitative analysis revealed that revisiting was one of the strongest causes of the PE slowdown, thus we were interested in finding out the ingredients of revisits. The major common modifications made during the revisits were term changes and structure level stylistic changes. What was particularly striking was that in some cases, a technical term was changed to another term, and then changed back to the previous one, and in some cases this was repeated multiple times swinging to and fro between two terms. This was a clear indication of the importance of standardisation of terminology and correct translation of them by MT systems. As for outlier cases, observation of screen capture video revealed the fundamental difference in editing processes between extremely slow PE cases and normal speed cases. While editing activities were constantly happening in normal cases, they were sporadic in slow cases with other activities, such as pausing, scrolling up and down, and referencing external information sources, occupying the majority of the PE time.

In addition, we strived to gain insight into professional or potential post-editors' thoughts and opinions by means of a questionnaire. A variety of suggestions were voiced as to the contents of possibly helpful PE training, with particular interests in MT output patterns and definitions of the anticipated quality of the target text, indicating that systematic information for understanding how to deal with MT and clear guidance as to the quality requirements are much desired. The biggest cause for inefficient PE in their everyday work environment, according to the participants, was the non-standardised elements in glossary and MT output, such as terminology, formats for certain elements, such as GUI items, use of symbols, spaces, and other special characters, and styles. This indicates the high awareness among post-editors and translators about standardisation in technical texts. MT output is expected to ensure standardisation to facilitate effective MT + PE workflow. In terms of MT and PE in general, a number of positive thoughts were expressed, specifically about the quality and potential of MT output, with some candid opinions about the limitations of current MT and the possible source of stress for post-editors.

9.3 Contribution to Knowledge

The present study contributes to the existing literature in the following ways:

- It observed real-life PE work by having professional post-editors and translators work for up to two full days in a practical working environment using real text data, standard translation and PE tools, and possibly one of the least invasive methods of observing the participants' behaviour. This made it possible to gather more realistic data than conducting a similar experiment in a controlled environment using tools specifically devised for the experiment. This makes the findings directly meaningful to the industry. In addition, as this was only possible by the financial and environmental support from the funding bodies, the present study highlighted the significant role of industry in academic research.
- It highlighted the difference between post-editing of MT output and editing of TM fuzzy matches from a qualitative perspective, which will make an informative topic in post-editor training materials.
- Through the questionnaire, it gathered opinions and thoughts from professional post-editors and translators about PE, a job for which the demand is increasingly high. Despite the rather small-scale survey, it picked up some candid and valuable opinions, which we hope will inform designing of PE training and guidelines as well as improvement of MT workflow.
- It spotlighted a much demanded language pair in IT localisation, English to Japanese; this was the first empirical study on PE of English-to-Japanese translation conducted in an authentic setting to the knowledge of the author of this thesis. Some, if not the majority, of the findings, however, cannot only be applicable to this specific language pair, but also be transferrable and helpful when addressing similar issues in other language pairs, especially in the case where English is the source language.
- It ventured to employ extensive statistical methods, which had generally been avoided in translation studies. Multiple regression analysis, which takes into account

a number of explanatory variables at a time, makes it possible not only to examine an effect of each of the various conditions while controlling for other variables, but also to quantify the combined effect of all conditions considered. This statistical technique, we believe, has a high potential in understanding and explaining human activities and behaviour in text processing, which are affected by a number of conditions in an intricate manner. The present study has shown the possibility of broader application of multiple regression analysis in the field of translation studies.

- We hope the knowledge obtained from our study generates benefits also for our industrial partner, Symantec. Our findings can provide a general view on various aspects of PE, especially for a much demanded but little studied language pair, English to Japanese. The results from our study provide a knowledge base for tuning-up Symantec's MT strategy, such as the reasonable targets for GTM scores and PE speed, balanced use of TM and MT, and ideas for future guidelines and training. Some of the findings may drive the motivation for reinforcing the existing methods and developing new methods, such as further utilising user dictionaries, developing an efficient method for handling of UI terms, and introducing MT quality estimation measures.

9.4 Limitations of the Study

9.4.1 MT system

The present study created a real-life PE setting for an observational experiment by taking advantage of the system that is currently in operation at Symantec, one of the funding bodies. While this was beneficial to a great extent, it also unavoidably limited the choice of system. Each MT system has its own strengths and weaknesses, thus some of the translation errors and resulting PE operations can be specific to the MT system employed, in the case of the present study, Systran. Also, as RBMT systems perform translation based on the defined grammar rules, they produce deterministic results, while SMT systems, which perform translation based on the models gleaned from the sample translations, produce probabilistic results. This difference may have a significant influence on the PE task; MT errors and necessary corrections probably have more similar patterns in the former case compared to the latter case. Thus the findings of the

present study may be more applicable for the PE of RBMT output than that of SMT output.

9.4.2 Language pair

The results of the present study were derived only from the observation of PE of English to Japanese translation. Although this does not necessarily limit the applicability of the findings of the present study to this specific language pair, it may have stronger implication to it compared to other language pairs. For instance, English to Japanese MT output often requires more extensive PE in general compared to English to European language pairs if we are to go by Symantec's experience. This might have affected both the amount of editing performed during PE and the time spent on the task. Also, as with any other languages, Japanese has specific writing conventions, such as polite form of sentence endings and ellipsis of pronouns, which may have resulted in some language-specific correction types.

9.4.3 Methodology

In an effort to reproduce a real-life PE environment while also gathering as much data as possible, we mainly made use of standard tools but with the aid of screen capture software and a specifically devised macro. Although they worked generally well, some participants, especially those who were using low spec PCs, experienced problems in functionality and performance. This might have slowed down the PE task to some extent, though such slowdown should not have affected the implication of the gathered data as each problem was consistent within each participant who experienced it.

In real life, the type and the size of the information provided to post-editors differ from one project to another; some projects are accompanied by extensive translation/PE guidelines, while others are carried out with limited guidelines. The present study used only a simplified set of guidelines. This was necessary since the allocated time was only two days, and it was preferable to have most of the time spent on actual PE rather than reading and assimilating the guidelines. However, some answers to the questionnaire revealed a certain level of confusion among some of the participants due to the lack of detailed guidelines for PE. This is a catch-22 situation; extensive and long guidelines

might equally have caused a problem by feeding the participants too much guidance to understand and follow in a short period of time.

9.5 Future Research

The present study focused on understanding the human PE effort and its problems, rather than developing and offering remedies to the problems. It is hoped, however, that the findings from the present study can form part of the building blocks from which future actions to solve the problems can take off.

The main interest of the study was to discover the elements that cause more PE effort independent of individual post-editors. However, during the course of an effort of finding such elements, we constantly encountered individual differences among participants. This was partly expected as each post-editor obviously had a different background, experience, methodology, ideology, and so on. Nevertheless, the scale of the individual difference was striking. Considering this fact, we would like to emphasise the importance of finding a balance between standardisation of the process and making the best use of personnel by respecting individual differences. We hope this, together with the findings of the present research, inform the future development of PE training and guidelines.

We discovered some potential for the complexity index, provided by Systran, in predicting the level of effort intensiveness for PE. This technology is related to a broader concept of ‘confidence index’,⁷⁹ a mechanism that predicts the quality of machine translation output and which can give post-editors some idea about how reliable the MT output is before reading the ST and MT output. Some of the findings from the present study as to the effect of ST characteristics, namely, the sentence structure, the document component parts, and the presence or absence of UI terms, could be considered as candidate variables to be incorporated in such an index to further refine the algorithm.

⁷⁹ There have been a number of related studies (Bernth 1999, Rojas & Aikawa 2006, Raybaud et al. 2009, Specia et al. 2009a, Specia et al. 2009b).

The study demonstrated the problematic nature of UI terms and technical terms in PE effort. This, in turn, suggests the possibility of further streamlining PE by taking the burden off post-editors' shoulders. We discovered that leaving UI terms in English could help reduce distraction for post-editors. Another possibility is tagging or adding annotations to terms that need not be edited, and incorporating such technology into the PE environment.⁸⁰

We also highlighted the different nature of editing of MT output compared to that of TM fuzzy matches. A comparison of these two translation-related activities will be an interesting topic in its own right. Moreover, many of the (potential) post-editors already have experience in translation, and in the case of the IT domain, most likely in editing of TM fuzzy matches too. Thus deeper understanding of such difference will greatly help to develop effective PE training courses.

We endeavoured to understand and quantify complex human activities by experimenting to apply multiple linear regression to the PE studies. This had never been done to our knowledge, which inevitably means it is not an established or agreed upon statistical analysis method in the field of PE study, although multiple regression itself is a proven technique to explain linear relationships and widely employed not only in non-linguistic fields of research, such as economics, political science, climatology, and pharmaceuticals, but is also very popular in applied linguistics research (Hatch & Lazaraton 1991). The application of multiple regression to our study was carefully chosen and planned after considering its requirements and assumptions, and its validity was examined by using various post-estimation tests. Nonetheless, further testing of the same technique in similar studies and exploring the potential benefit of applying other statistical techniques, non-linear models for instance, may well be merited in the future.

⁸⁰ A similar idea has been suggested by Roturier & Lehmann (2009).

REFERENCES

- Abekawa, T. & Kageura, K. (2008a) Constructing a corpus that indicates patterns of modification between draft and final translations by human translators. In: *Proceedings of the Sixth international conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, pp. 2000-2005.
- Abekawa, T. & Kageura, K. (2008b) What prompts translators to modify draft translations? An analysis of basic modification patterns for use in the automatic notification of awkwardly translated text. In: *Proceedings of the Third International Joint Conference on Natural Language Processing*, Hyderabad, India, pp. 241-248.
- Abekawa, T. & Okumura, M. (2006) Japanese Dependency Parsing Using Co-occurrence Information and a Combination of Case Elements. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (ACL/Coling 2006)*, pp. 833-840.
- Aikawa, T., Schwartz, L., King, R., Mo, C.O. & Lozano, C. (2007) Impact of Controlled Language on Translation Quality and Post-editing in a Statistical Machine Translation Environment. In: *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 1-7.
- Aizawa, T., Ehara, T., Uratani, N., Tanaka, H., Kato, N., Nakase, S., Aruga, N. & Matsuda, T. (1990) A machine translation system for foreign news in satellite broadcasting. In: H. Karlgren (ed) *Proceedings of the 13th International Conference on Computational Linguistics (Coling 90)*, Helsinki, pp. 308-310.
- Allen, J. (2003) Post-editing. In: H. Somers (ed) *Computers and Translation: A Translator's Guide*. John Benjamins Publishing, Amsterdam, pp. 297-317.
- Allen, J. (2001) Postediting: an integrated part of a translation software program. *Language International*, 13, 2, 26-29.
- Allen, J. & Hogan, C. (2000) Toward the Development of a Postediting Module for Raw Machine Translation Output: a Controlled Language Perspective. In: *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000)*, Seattle, Washington, pp. 62-71.
- ALPAC (1966) *LANGUAGE AND MACHINES: COMPUTERS IN TRANSLATION AND LINGUISTICS*. Division of Behavioral Sciences, National Academy of Sciences, National Research Council.
- Arapakis, I., Jose, J.M. & Gray, P.D. (2008) Affective feedback: an investigation into the role of emotions in the information seeking process. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 395-402.
- Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L. & Sadler, L. (1993) *Machine Translation: An Introductory Guide*. NCC Blackwell, London.
- Austermühl, F. (2001) *Electronic tools for translators*. St. Jerome, Manchester; Northampton, MA.

- Bernth, A. (1999) A Confidence Index for Machine Translation. In: *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England, pp. 120-127.
- Bernth, A. & Gdaniec, C. (2001) MTranslatability. *Machine Translation*, 16, 3, 175-218.
- Biber, D. (1988) *Variation across speech and writing*. Cambridge University Press, Cambridge [England]; New York.
- Boitet, C., Bey, Y., Tomokio, M., Cao, W. & Blanchon, H. (2006) IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations. In: *Proceedings of International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation [IWSLT 2006]*, Kyoto, Japan, pp. 8-15.
- Bond, F., Ogura, K. & Ikehara, S. (1994) Countability and number in Japanese to English machine translation. In: *Proceedings of the 15th conference on Computational linguistics*, Kyoto, Japan, pp. 32-38.
- Boslaugh, S. & Watters, P.A. (2008) *Statistics in a nutshell*. 1st edn, O'Reilly, California.
- Bowker, L. (2002) *Computer-aided translation technology*. University of Ottawa Press, Ottawa.
- Byrne, J. (2006) *Technical translation : usability strategies for translating technical documentation*. Springer, Dordrecht.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2008) Further Meta-Evaluation of Machine Translation. In: *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, USA, Association for Computational Linguistics, pp. 70-106.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2007) (Meta-) Evaluation of Machine Translation. In: *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, pp. 136-158.
- Campbell, S. (2000) Choice Network Analysis in Translation Research. In: M. Olohan (ed) *Intercultural Faultliness: Research Models in Translation Studies: Textual and Cognitive Aspects*. St. Jerome, Manchester, pp. 29-42.
- Carl, M., Jakobsen, A.L. & Jensen, K.T.H. (2008) Modelling Human Translator Behaviour with User-Activity Data. In: *Proceedings of the 12th Annual conference of the European Association for Machine Translation*, Hamburg, Germany, pp. 21-26.
- Christians, C.G. (2003) Ethics and Politics in Qualitative Research. In: N.K. Denzin & Y.S. Lincoln (eds) *The Landscape of Qualitative Research: Theories and Issues*. SAGE Publications, Thousand Oaks, London, New Delhi, pp. 208-243.
- Creswell, J.W. & Plano Clark, P.L. (2007) *Designing and conducting mixed methods research*. SAGE Publications, Thousand Oaks, Calif.
- de Almeida, G. & O'Brien, S. (2010) Analysing Post-Editing Performance: Correlations with Years of Translation Experience. In: V. Hansen & F. Yvon (eds) *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint-Raphaël, France.

- De Roeck, A., Sarkar, A. & Garthwaite, P.H. (2005) Even very frequent function words do not distribute homogeneously. In: *Proceedings of 2005 International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovetz, Bulgaria.
- DePalma, D.A. & Kelly, N. (2009) *The Business Case for Machine Translation: How Organizations Justify and Adopt Automated Translation*. Common Sense Advisory, Inc.
- Désilets, A., Brunette, L., Melançon, C. & Patenaude, G. (2008) Reliable innovation: a tecchie's travels in the land of translators. In: *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawai'i, pp. 339-345.
- Dillinger, M. & Lommel, A. (2004) *LISA Best Practice Guides: Implementing Machine Translation*.
- Directorate-General for Translation (2008) *Translation Tools and Workflow*, European Commission, Brussels.
- Doherty, S. & O'Brien, S. (2009) Can MT output be evaluated through eye tracking? In: *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, pp. 214-221.
- Doherty, S., O'Brien, S. & Carl, M. (2010) Eye tracking as an MT evaluation technique. *Machine Translation*, 24, 1-13.
- Douglas, S. & Hurst, M. (1996) Controlled Language Support for Perkins Approved Clear English (PACE). In: *Proceedings of the First Controlled Language Application Workshop (CLAW 1996)*, Leuven, Belgium, Centre for Computational Linguistics, pp. 93-105.
- Dugast, L., Senellart, J. & Koehn, P. (2007) Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In: *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, pp. 220-223.
- Echizen-ya, H., Ehara, T., Shimohata, S., Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T. & Kando, N. (2009) Meta-evaluation of automatic evaluation methods for machine translation using patent translation data in NTCIR-7. In: *Proceedings of MT Summit XII: Third Workshop on Patent Translation*, Ottawa, Ontario, Canada, pp. 9-16.
- Ehara, T. (2007a) A Proposal for A New Automatic Machine Translation Evaluation Metrics: NMG. [新しい機械翻訳自動評価基準 NMG の提案] *Japio 2007 YEARBOOK*, 238-241.
- Ehara, T. (2007b) Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In: *Proceedings of MT Summit XI: Workshop on patent translation*, Copenhagen, Denmark, pp. 13-18.
- Fiederer, R. & O'Brien, S. (2009) Quality and machine translation: a realistic objective? *Journal of Specialised Translation*, 11, 52-74.
- Flanagan, M. (1997) MT Today: Emerging Roles, New Successes. *Machine Translation*, 12, 25-27.
- Frey, L.R., Botan, C.H. & Kreps, G.L. (1991) *Investigating Communication - An Introduction to Research Methods*. Prentice Hall, Englewood Cliffs, H. J.

- Fujii, A., Utiyama, M., Yamamoto, M. & Utsuro, T. (2009) Exploiting Patent Information for the Evaluation of Machine Translation. In: *Proceedings of MT Summit XII: Third Workshop on Patent Translation*, Ottawa, Ontario, Canada, pp. 9-16.
- Funaki, S. (1993) Multi-lingual machine translation (MMT) project. In: *Proceedings of MT Summit IV: International Cooperation for Global Communication*, Kobe, Japan, pp. 73-78.
- Gamon, M., Aue, A. & Smets, M. (2005) Sentence-level MT evaluation without reference translations: beyond language modeling. In: *Proceedings of the 10th Annual conference of the European Association for Machine Translation*, Budapest, pp. 103-111.
- Garcia, I. (2010) Is machine translation ready yet? *Target*, 22, 1, 7-21.
- Garcia, I. (2007) Power shifts in web-based translation memory. *Machine Translation*, 21, 1, 55-68.
- González-Rubio, J., Ortiz-Martínez, D. & Casacuberta, F. (2010) Balancing user effort and translation error in interactive machine translation via confidence measures. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 173-177.
- Groves, D. & Schmidtke, D. (2009) Identification and analysis of post-editing patterns for MT. In: *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, pp. 429-436.
- Guerberof, A.A. (2008) Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus: The International Journal of Localisation*, 7, 1, 11-21.
- Gustavii, E. (2005) Target language preposition selection - an experiment with transformation based learning and aligned bilingual data. In: *Proceedings of the 10th Annual conference of the European Association for Machine Translation*, Budapest, pp. 112-118.
- Hague, P.N. (1993) *Questionnaire Design (Market Research)*. Kogan Page Ltd.
- Hatch, E.M. & Lazaraton, A. (1991) *The Research Manual: Design and Statistics for Applied Linguistics*. Heinle and Heinle, Boston, MA.
- Hayashi, O., Ikegami, A. & Ando, C. (eds) (2004) *Dictionary for understanding Japanese[日本語文法がわかる事典]*, Tokyodo Publishing, Tokyo, Japan.
- Healey, P.G.T., Vogel, C. & Eshgi, A. (2007) Group Dialects in an Online Community. In: R.V. Arnstein L. (ed) *Proceedings of DECALOG 2007, The 10th Workshop on the Semantics and Pragmatics of Dialogue*, Università di Trento, Italy, pp. 141-147.
- Hughes, M.A. & Hayhoe, G.F. (2007) *A Research Primer for Technical Communication: Methods, Exemplars, and Analyses*. 1st edn, Routledge.
- Hurst, S. (2008) Trends in automated translation in today's global business. In: *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawai'i, pp. 394-396.
- Hutchins, J. (2003) Commercial systems. In: H. Somers (ed) *Computers and Translation: A Translator's Guide*. John Benjamins Publishing, Amsterdam, pp. 161-174.

- Isahara, H. (1995) JEIDA's test-sets for quality evaluation of MT systems – technical evaluation from the developer's point of view. In: *Proceedings of MT Summit V*, Luxembourg.
- Isahara, H., Kurohashi, S., Tsujii, J., Uchimoto, K., Nakagawa, H., Kaji, H. & Kikuchi, S. (2007) Development of a Japanese-Chinese machine translation system. In: *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 263-267.
- Ishisaka, T., Yamamoto, K., Utiyama, M. & Sumita, E. (2009) Development of a Japanese-English software manual parallel corpus. In: *Proceedings of MT Summit XII*, Ottawa, Canada, pp. 254-259.
- Itagaki, M., Kosaka, T. & Ohno, Y. (1999) *The Handbook of Software Localization*. [ソフトウェアローカリゼーション実践ハンドブック ~海外ソフトウェア開発の現場から~] Soft Research Center Inc., Tokyo.
- Iwabuchi, T. (2000) *Japanese Grammar*. [日本語文法] Hakuteisha, Tokyo, Japan.
- Iwatate, M., Asahara, M. & Matsumoto, Y. (2008) Japanese Dependency Parsing Using a Tournament Model. In: *Proceedings of The 22nd International Conference on Computational Linguistics (Coling 2008)*, Coling 2008 Organizing Committee, pp. 361-368.
- Jakobsen, A.L. (1998) Logging time delay in translation. *LSP Texts and the Process of Translation*, 73-101.
- Japan Translation Federation (2009) *White Paper: The Third Research Report on Translation Industry*. [平成 20 年度 翻訳白書 : 第 3 回 翻訳業界調査報告書] Japan Translation Federation.
- Johansson, S., Leech, G.N. & Goodluck, H. (1978) *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*, Department of English, University of Oslo, Oslo.
- Karamanis, N., Luz, S. & Doherty, G. (2010) Translation practice in the workplace and machine translation. In: V. Hansen & F. Yvon (eds) *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint-Raphaël, France, pp. 8-15.
- Kilgarriff, A. (2001) COMPARING CORPORA. *International Journal of Corpus Linguistics*, 6, 1, 97-133.
- Kilgarriff, A. (1997) Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. In: *Proceedings of the 5th ACL workshop on very large corpora*, Beijing and Hong Kong, pp. 231-245.
- Kiuchi, T. & Kaihara, S. (1991) Quantitative evaluation of English-Japanese machine translation of medical literature. *Methods of Information in Medicine*, 3, 199-205.
- Knight, K. & Chander, I. (1994) Automated Postediting of Documents. In: *Proceedings of the 12th National conference of the American Association for Artificial Intelligence (AAAI 1994)*, Seattle, Washington, USA.
- Koehn, P. (2009) A process study of computer-aided translation. *Machine Translation*, 23, 4, 241-263.
- Kohl, J.R. (2008) *The Global English Style Guide: Writing Clear, Translatable Documentation for a Global Market*. SAS Institute Inc., Cary, NC.

- Kondo, M. & Wakabayashi, J. (2009) Japanese tradition. In: M. Baker & G. Saldanha (eds) *Routledge Encyclopedia of Translation Studies*. 2nd edn, Routledge, London & New York, pp. 468-476.
- Kosaka, T. (2002) *Translation of Scientific and Technical Texts*. [<英 日> 技術翻訳の A to Z] Kenkyusha, Tokyo.
- Krings, H.P. (2001) *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. The Kent State University Press, Kent, Ohio.
- Kudo, T. & Matsumoto, Y. (2002) Japanese dependency analysis using cascaded chunking. In: *Proceedings of the 6th conference on Natural language learning - Volume 20, International Conference On Computational Linguistics*, Association for Computational Linguistics, pp. 1-7.
- Kumano, A. (2008) Challenges in Machine Translation of Japanese to English Patent Translation. [特許明細書の日英機械翻訳における課題] *Japio 2008 YEARBOOK*, 88-93.
- Kumano, A., Yasuhara, H. & Watanabe, T. (2009) “Technical Japanese” Platform; Experiment of Paraphrasing Patent Text. [産業日本語の構想と特許文の言い換え実験 (日本語処理・文法)] *IPSJ SIG Notes*, 2009(36), 15-20.
- Lagoudaki, E. (2006) *Translation Memories Survey 2006, Translation Memory systems: Enlightening users' perspective*. Imperial College London.
- Laurian, A.M. (1984) Machine Translation: What Type of Post-Editing on What Type of documents for What Type of Users. In: *Proceedings of the 10th International Conference on Computational Linguistics (Coling 84)*, Stanford University, pp. 236-238.
- LDC (2005) *Linguistic Data Annotation Specification: Assessment of fluency and adequacy in translations*. Report number: Revision 1.5.
- Leech, G. (2006) *A Glossary of English Grammar*. Edinburgh University Press Ltd, Edinburgh.
- Lin, C.-. & Och, F.J. (2004) ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In: *Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004)*, University of Geneva, Switzerland, pp. 7.
- LISA *LISA QA Model Version 3.1*. LISA.
- Llitjós, A.F., Carbonell, J. & Lavie, A. (2007) Improving Transfer-Based MT Systems with Automatic Refinements. In: *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 183-190.
- Loizides, F. & Buchanan, G. (2009) An Empirical Study of User Navigation during Document Triage. In: M.e.a. Agosti (ed) *Research and Advanced Technology for Digital Libraries*. Springer, Berlin / Heidelberg, pp. 138-149.
- Makita, M., Higuchi, S., Fujii, A. & Ishikawa, T. (2003) A system for Japanese/English/Korean multilingual patent retrieval. In: *Proceedings of MT Summit IX*, New Orleans, USA, pp. 475-478.

- Mamidi, R. (2004) Disambiguating Prepositions for Machine Translation using Lexical Semantic Resources. In: *Proceedings of National Seminar on Theoretical and Applied Aspects of Lexical Semantics*, Osmania University at Hyderabad.
- Markel, M. (2003) *Technical Communication*. 6th edn, Bedford/St. Martin's, Boston & New York.
- McElhaney, T. & Vasconcellos, M. (1988) *The Translator and the Postediting Experience*, State University of New York at Binghamton (SUNY).
- McEnery, T. & Wilson, A. (1996) *Corpus linguistics*. Edinburgh University Press, Edinburgh.
- Melamed, I.D., Green, R. & Turian, J.P. (2003) Precision and Recall of Machine Translation. In: *Proceedings of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 61-63.
- Mitamura, T. (1999) Controlled Language for Multilingual Machine Translation. In: *Proceedings of MT Summit VII*, Singapore, pp. 46-52.
- Mitamura, T. & Nyberg, E. (1995) Controlled English for Knowledge-Based MT: Experience with the KANT System. In: *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, pp. 158-172.
- Mossop, B. (2007) Empirical studies of revision: what we know and need to know. *The Journal of Specialised Translation*, 8, 5-20.
- Murata, K., Yamada, K. & Sakuma, I. (2007) *Methodology for Social Psychology*. [社会心理学研究法] Fukumura Shuppan Inc.
- Nagao, M. (1984) A framework of a mechanical translation between Japanese and English by analogy principle. In: *Proceedings of the international NATO symposium on Artificial and human intelligence, October 1981*. Lyon, France, Elsevier North-Holland, Inc. pp. 173-180.
- Nagao, M., Tsuji, J., Mitamura, K., Hirakawa, H. & Ku, M. (1980) A machine translation from Japanese into English – another perspective of MT systems. In: *Proceedings of the 8th International Conference on Computational Linguistics (Coling 80)*, Tokyo, Japan, pp. 414-423.
- Nagao, M., Tsujii, J., Yada, K. & Kakimoto, T. (1982) An English Japanese machine translation system of the titles of scientific and engineering papers. In: J. Horecký (ed) *Proceedings of the Ninth International Conference on Computational Linguistics (Coling 82)*, Prague, North-Holland Publishing Company, pp. 245-252.
- Nakamura, J. (1993) Statistical Methods and Large Corpora. In: M. Baker, G. Francis & E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair*. John Benjamins Publishing Comp, Philadelphia, pp. 293-311.
- National Institute of Standards and Technology Information Access Division / Speech Group And Linguistic Data Consortium (2007) *Post Editing Guidelines For GALE Machine Translation Evaluation*. Report number: Version 3.0.2.
- National Research Council (2002) *Scientific Research in Education*. National Academy Press, Washington, DC.

- Neumann, C. (2005) A human-aided machine translation system for Japanese-English patent translation. In: *Proceedings of MT Summit X: Workshop on Patent Translation*, Phuket, Thailand, pp. 1-7.
- Newton, J. (1992) The Perkins experience. In: J. Newton (ed) *Computers and Translation: A practical appraisal*. Routledge, London, pp. 46-57.
- Nomura, H. (1993) Evaluation method of machine translation: from the viewpoint of natural language processing. Strategy for machine translation evaluation. In: *Proceedings of MT Summit IV: International Cooperation for Global Communication*, Kobe, Japan, pp. 205-207.
- Nomura, H. & Isahara, H. (1992) Evaluation Surveys: The JEIDA Methodology and Survey. In: *Proceedings of MT Evaluation: Basis for Future Directions, a workshop sponsored by the National Science Foundation*, San Diego, CA, pp. 11-12.
- Nyberg, E. & Mitamura, T. (1996) Controlled Language and Knowledge-Based Machine Translation: Principles and Practice. In: *Proceedings of the First Controlled Language Application Workshop (CLAW 1996)*, Leuven, Belgium, Centre for Computational Linguistics, pp. 74-83.
- Nyberg, E., Mitamura, T. & Huijsen, W.O. (2003) Controlled language for authoring and translation. In: H. Somers (ed) *Computers and Translation: A Translator's Guide*. John Benjamins Publishing, Amsterdam, pp. 245-281.
- Oakes, M.P. (1998) *Statistics for corpus linguistics*. Edinburgh University Press, Edinburgh.
- O'Brien, S. (2007) An Empirical Investigation of Temporal and Technical Post-Editing Effort. *Translation and Interpreting Studies*, 2, 1, 83-136.
- O'Brien, S. (2006a) Eye-tracking and Translation Memory Matches. *Perspectives Studies in Translatology*, 14, 3, 185-205.
- O'Brien, S. (2006b) *Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*, Dublin City University.
- O'Brien, S. (2006c) Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. *Across Languages and Cultures*, 7, 1, 1-21.
- O'Brien, S. (2005) Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation*, 19, 37-58.
- O'Brien, S. (2002) Teaching post-editing: a proposal for course content. In: *Proceedings of the sixth EAMT Workshop "Teaching machine translation"*, UMIST, Manchester, England, pp. 99-106.
- O'Brien, S. & Roturier, J. (2007) How Portable are Controlled Language Rules?: A Comparison of Two Empirical MT Studies. In: *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 345-352.
- Oppenheim, A.N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. 2nd edn, Pinter Pub Ltd, London.
- Oshio, T. (1980) Applied testing of HICATS/JE for Japanese patent abstracts. In: M. Nagao (ed) *Machine Translation Summit*. Ohmsha, Tokyo, pp. 140-144.

- Papineni, K., Roukos, S., Ward, T. & Zhu, W. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 311-318.
- Parsons, K., McCormac, A. & Butavicius, M. (2009) *Human Dimensions of Corpora Comparison: An Analysis of Kilgarriff's (2001) Approach*. Command, Control, Communications and Intelligence Division, DSTO Defence Science and Technology Organisation. Report number: DSTO-TR-2290.
- Plitt, M. & Masselot, F. (2010) A productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93, 7-16.
- Povlsen, C. & Bech, A. (2001) Ape: reducing the monkey business in post-editing by automating the task intelligently. In: *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, pp. 283-286.
- Ramirez, L.P. & Haller, J. (2005) Controlled Language and the Implementation of Machine Translation for Technical Documentation. In: *Proceedings of ASLIB Conference, Translation and the Computer 27*, London.
- Rasinger, S.M. (2008) *Quantitative Research in Linguistics*. Continuum International Publishing Group, London, New York.
- Raybaud, S., Lavecchia, C., Langlois, D. & Smaili, K. (2009) Word- and sentence-level confidence measures for machine translation. In: L. Màrquez & H. Somers (eds) *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Universitat Politècnica de Catalunya, Barcelona, Spain, pp. 104-111.
- Rojas, D.M. & Aikawa, T. (2006) Predicting MT quality as a function of the source language. In: *Proceedings of the Fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, pp. 2534-2537.
- Roturier, J. (2009) Deploying novel MT technology to raise the bar for quality: a review of key advantages and challenges. In: *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, pp. 1-8.
- Roturier, J. (2006) *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users*, Dublin City University.
- Roturier, J. & Lehmann, S. (2009) How to Treat GUI Options in IT Technical Texts for Authoring and Machine Translation. *The Journal of Internationalisation and Localisation*, 1, 40-59.
- Rychtycky, N. (2006) Standard Language at Ford Motor Company: A Case Study in Controlled Language Development and Deployment. In: *Proceedings of the Fourth Controlled Language Application Workshop (CLAW 2006)*, Boston Marriott, pp. 10.
- Schäfer, F. (2003) MT post-editing: how to shed light on the “unknown task”. Experiences at SAP. In: *Proceedings of Controlled language translation, EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, pp. 133-140.
- Schilperoord, J. (1996) *It's About Time: Temporal Aspects of Cognitive Processes in Text Production*. Rodopi, Amsterdam.
- SDL Research (2008) *Trends in automated translation in today's global business*. SDL.

- Shih, C.Y. (2006) Revision from translator's point of view: an interview study. *Target: International journal of translation studies*, 18, 2, 295-312.
- Simard, M., Goutte, C. & Isabelle, P. (2007a) Statistical Phrase-based Post-editing. In: *Proceedings of NAACL-HLT-2007 Human Language Technology: the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, pp. 508-515.
- Simard, M., Ueffing, N., Isabelle, P. & Kuhn, R. (2007b) Rule-based translation with statistical phrase-based post-editing. In: *Proceedings of ACL 2007: The Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 203-206.
- Snoover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.
- Song, S. & Bond, F. (2009) Online search interface for the Sejong Korean-Japanese bilingual corpus and auto-interpretation of phrase alignment. In: *Proceedings of ACL-IJCNLP 2009: Third Linguistic Annotation Workshop (LAW III)*, Suntec, Singapore, pp. 146-149.
- Soricut, R. & Echihiabi, A. (2010) TrustRank: inducing trust in automatic translations via ranking. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 612-621.
- Specia, L., Cancedda, N., Dymetman, M., Turchi, M. & Cristianini, N. (2009a) Estimating the sentence-level quality of machine translation systems. In: L. Màrquez & H. Somers (eds) *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Universitat Politècnica de Catalunya, Barcelona, Spain, pp. 28-35.
- Specia, L., Saunders, C., Turchi, M., Wang, Z. & Shawe-Taylor, J. (2009b) Improving the confidence of machine translation quality estimates. In: *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, pp. 136-143.
- Sumita, E., Shimizu, T. & Nakamura, S. (2007) NICT-ATR speech-to-speech translation system. In: *Proceedings of ACL 2007*, Prague, Czech Republic, pp. 25-28.
- SYSTRAN *SYSTRAN 6 Desktop User Guide*. SYSTRAN.
- Takahashi, S., Wada, H., Tadenuma, R. & Watanabe, S. (2003) English-Japanese Machine Translation. In: S. Nirenburg, H. Somers & Y. Wilks (eds) *Readings in Machine Translation*. MIT Press, pp. 193-199.
- Takahashi, T. & You, K. (1990) *Planning and Analyses of Questionnaire Survey*. [質問紙調査の計画と解析] Bunka Publishing Bureau, Japan.
- Tatsumi, M. (2009) Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. In: *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, pp. 332-339.
- Tatsumi, M. & Sun, Y. (2008) A Comparison of Statistical Post-Editing on Chinese and Japanese. *Localisation Focus: The International Journal of Localisation*, 7, 1, 22-33.
- TAUS (2010) *Postediting in Practice*. TAUS BV.

- TAUS (2009) *Increasing leveraging from shared industry data*. Translation Automation User Society.
- Temnikova, I. (2010) Cognitive evaluation approach for controlled language post-editing experiment. In: *Proceedings of the Seventh international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, pp. 3485-3490.
- Temnikova, I. & Orasan, C. (2009) Post-editing experiments with MT for a controlled language. In: *Proceedings of ISMTCL: International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages, and their application to emergencies and safety critical domains*, Centre Tesnière, University of Franche-Comté, Besançon, France, pp. 249-255.
- Tort, F., Blondel, F.M. & Bruillard, E. (2009) From error detection to behaviour observation: first results from screen capture analysis. In: *Proceedings of European Spreadsheet Risks Int. Grp.*, pp. 119-132.
- Turian, J.P., Shen, L. & Melamed, I.D. (2003) Evaluation of Machine Translation and its Evaluation. In: *Proceedings of MT Summit IX*, New Orleans, USA, pp. 386-393.
- Txabarriaga, R., Kelly, N. & Stewart, R.G. (2009) *The European Translation Market: Main Sectors and Drivers for Language Services in Europe*. Common Sense Advisory, Inc.
- Uchimoto, K., Sekine, S. & Isahara, H. (1998) Japanese Dependency Structure Analysis based on Maximum Entropy Models. [*ME による日本語係り受け解析*] *IPSJ SIG Notes*, 98(99), 31-38.
- Underwood, N.L. & Jongejan, B. (2001) Translatability Checker: A Tool to Help Decide Whether to Use MT. In: *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, pp. 363-368.
- Usui, T., Sakamoto, K., Sawada, S., Yoshimura, K. & Kaji, H. (1986) Post-editing Environment Supported by HICATS/JE. [*HICATS/JE のポストエディット支援環境*] In: *Proceedings of the 33th annual conference of Information Processing Society of Japan*, pp. 1759-1760.
- Van Gijssel, S. & Vogel, C. (2003) Inducing a cline from corpora of political manifestos. In: *Proceedings of the 1st international symposium on Information and communication technologies, WORKSHOP SESSION: Invited workshop on conceptual information retrieval and clustering of documents table of contents*, Dublin, Ireland, Trinity College Dublin, pp. 297-303.
- Vasconcellos, M. (1987) Post-editing on-screen: machine translation from Spanish into English. In: C. Picken (ed) *Proceedings of Translating and the Computer 8: A Profession on the Move*, Aslib, London, pp. 133-146.
- Wagner, E. (1987) Post-Editing: Practical Considerations. In: C. Picken (ed) *Proceedings of ITI Conference. 1: The Business of Translating and Interpreting*, Aslib, pp. 71-78.
- Wagner, E. (1985) Rapid Post-Editing of Systran. In: V. Lawson (ed) *Tools for the trade : translating and the computer 5*. Aslib, London, pp. 199-214.
- Watanabe, T. (2010) Technical Japanese as an Innovation Infrastructure: Development of Patent Technical Japanese and Technical Japanese Platform. [*イノベーション*]

インフラとしての産業日本語: 特許版産業日本語と産業日本語プラットフォームの開発について] *Japio 2010 YEARBOOK*, 160-165.

- Wendt, C. & Lewis, W. (2009) Pushing the quality of a customized SMT system using shared training data. In: *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada.
- White, J. (2003) How to Evaluate Machine Translation. In: H. Somers (ed) *Computers and Translation: A Translator's Guide*. John Benjamins Publishing, Amsterdam, pp. 201-232.
- Williams, G. (2002) In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, 7, 1, 43-64.
- Woods, A., Fletcher, P. & Hughes, A. (1986) *Statistics in language studies*. Cambridge University Press, Cambridge.
- Wooldridge, J.M. (2006) *Introductory Econometrics: A Modern Approach*. 3rd edn, Thomson South-Western, Ohio.
- Wu, X., Cardey, S. & Greenfield, P. (2006) Some Problems of Prepositional Phrases in Machine Translation. In: *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, pp. 593-603.
- Yamada, M. (2010) Applying “machine translation plus post-editing” to a case of English-to-Japanese translation. [「機械翻訳は使えるのか - Post-editing の観点から検証する - 」] In: *Proceedings of the 11th Annual Conference of the Japan Association for Interpreting and Translation Studies*, Tokyo, Japan.
- Yamamoto, K. (1999) Proofreading Generated Outputs: Automated Rule Acquisition and Application to Japanese-Chinese Machine Translation. In: *Proceedings of ICCPOL-99*, Tokushima, Japan, pp. 87-92.
- Yokoyama, S. & Kennendai, S. (2007) Error correcting system for analysis of Japanese patent sentences. In: *Proceedings of MT Summit XI: Workshop on patent translation*, Copenhagen, Denmark, pp. 24-27.
- Yoshimi, T. (2001) Improvement of Translation quality of English Newspaper Headlines by Automatic Pre-editing. *Machine Translation*, 16, 233-250.
- Yoshimi, T., Sata, I. & Fukumochi, Y. (2000) Automatic Preediting of English Sentences for a Robust English-to-Japanese MT System. [頑健な英日機械翻訳システム実現のための原文自動前編集] *Natural Language Processing [自然言語処理]*, 7, 4, 99-117.
- Yoshimura, K. (2000) *Basics of Natural Language Processing*. [自然言語処理の基礎] Saiensu-sha, Tokyo, Japan.

APPENDICES

Appendix A PROJECT MANUAL

This appendix includes the actual project manual used for the project (in Japanese), and its English translation (translated by the author of this thesis).

Oak プロジェクトマニュアル

Oak プロジェクトにご参加いただくポストエディタの方へ

2009 年 7 月

Rev.1.4

今回はポストエディティング調査プロジェクトにご協力いただき、ありがとうございます。
この文書は以下の3つのセクションで構成されています。

- I. プロジェクトの概要
- II. 作業手順書
- III. ポストエディティングガイドライン

すべてお読みになってからポストエディティング作業を開始してください。
ご不明な点、説明どおりに動作しない点がある場合は、ポストエディティング作業を続行せず、
ご質問ください。

I. プロジェクトの概要

< 目的 >

このプロジェクトの最終的な目的は、ポストエディティング (PE) タスクの負担を特に増加させている原因 (原文の問題点や機械翻訳のエラーなど) を詳しく突き止めることです。そのために、セグメントごとの PE 作業時間を計測し、作業に時間がかかってしまうセグメントの特徴を分析します。

< データと個人情報の取り扱いについて >

時間の計測や画面の録画は、原文や訳文の問題点を探ることが目的であり、ポストエディタの方の能力を測るためではありません。この調査で測定されたデータは、この調査の目的以外で使用されることはなく、またポストエディタの方について個人が特定できるような情報を公開することはありません。

< テキストについて >

PE していただくのは、Symantec の製品のマニュアルから抜粋したテキストです。

< 作業ペースについて >

PE していただくのは、32 ファイル、英文でのワードカウントで 5,029 ワード分です。およそ一日程度を目安と考えていますが、時間制限は設けていません。末尾のガイドラインに沿って、通常の業務と同様のペースで行ってください。作業中はなるべく一定のペースを保つようにしてください。途中で何度休憩していただいてもかまいません。

< 作業手順について >

PE には TagEditor の標準的なキー操作を使用していただきますが、正確なデータ収集のため、若干追加の手順が必要になります。次のセクション『作業手順書』を参照の上、事前にテスト用のファイル (_Test_kit フォルダ内にあります) を使用して手順や動作を確認してから開始してください。説明どおりに動作しない場合は、ポストエディティングを開始せず、立見みどりまでご質問ください。

II. 作業手順書

1. Oak_kit のフォルダ構造とファイル

はじめにお読みください_Oak.pdf (このファイル)

アンケート.doc (プロジェクト終了後に使用します)

_Test_kit [テストおよび練習用のファイル]

_Pause_Restart.xml (作業を一時中断する際に使用します)

Test_kit.* (Trados 7 用の TM)

Test_kit65.* (Trados 6.5 用の TM)

v17219425_ja_jp_17327647.xml (PE 練習用ファイル)

PE [PE 時に使用するファイル]

*.xml (PE 対象のファイル)

_Pause_Restart.xml (ファイル内で作業を一時中断する際に使用します)

Settings [各種設定ファイル]

Oak_Key.exe (作業前にダブルクリックして起動します)

Ini [設定ファイル]

TM [翻訳メモリファイル]

OAK.* (Trados 7 用の TM)

OAK65.* (Trados 6.5 用の TM)

フォルダ構造は変更しないでください。

作業終了後は、フォルダ構造ごとすべてのファイルを納品してください。

2. ツールの設定

➤ TRADOS Workbench の設定

作業を開始する前に、TRADOS Translator's Workbench を起動して、以下の操作を実行してください。

● ID の指定

1. すでに何らかの TM が開いている場合は閉じます。
2. [Settings (環境)] メニューから [User ID (ユーザー ID)] を選択します。
3. テキストボックスに「PExx」と入力して、[OK] をクリックします。(x にはお知らせする ID 番号を入れます。)

● TM のオープン

1. [ファイル] メニューから [開く] を選択し、TM フォルダにある「OAK.tmw」(Trados 6.5 の場合は「OAK65.tmw」) を選択して、[開く] をクリックします。

● TM オプションの指定

1. [オプション] メニューから [翻訳メモリオプション] を選択します。
2. [全般] タブの一致精度最小値を「70%」に設定します。
3. [属性およびテキストフィールドの更新] のオプションとして両方とも [結合] を選びます。
4. [ペナルティ] タブを次のように設定します。機械翻訳によるペナルティは 26% に設定してください。

異なる書式によるペナルティ %(F):	1	<input type="range"/>
属性とテキスト フィールドの不一致によるペナルティ %(I):	2	<input type="range"/>
固定要素の不一致によるペナルティ %(P):	2	<input type="range"/>
整合ペナルティ %(A):	3	<input type="range"/>
機械翻訳によるペナルティ %(M):	26	<input type="range"/>
複数の訳文によるペナルティ %(L):	0	<input type="range"/>

☐ 原文のタグが異なるときはペナルティを課す

5. [ツール] タブの [タグ設定] ボタンをクリックして、[開く] をクリックし、Oak_kit\Setting\ini\docbookx.ini を開きます。
6. このファイルを [既定値] にして [OK] を 2 回クリックします。

これで Workbench の設定が完了しました。

TagEditor でファイルを開く際には、ここで設定したファイル(「Symbook」と表示されます)を使用してください。

➤ マクロの起動

TagEditor に「Oak_Key」というマクロを組み合わせで使用します。これにより、標準的な TagEditor のショートカットキーを使用していただけで、調査データの記録が可能になります。

- a. Settings フォルダにある **Oak_Key.exe** をダブルクリックします。
- b. [H]アイコンが Windows のタスクトレイに入ったことを確認してください。



これでマクロの設定が完了しました。

- ☒ ポストエディティング作業完了後に、[H]アイコンを右クリックして[Exit]を選択し、マクロを終了してください。

➤ BB FlashBack 2 Express の起動

BB FlashBack は、画面上の動作を記録するためのプログラムです。ポストエディティング作業の内容を分析するため、このプログラムを使用して作業内容を記録します。

1. [スタート]メニューの[プログラム]から[Blueberry Software]→[BB FlashBack 2 Express (Japanese)]→[BB FlashBack 2 Express プレーヤー]の順に選択します。
2. タスクトレイの下矢印をクリックし、[クイックレコード]を選択します。



3. [OK]をクリックします。
4. これで録画が開始されます。
5. 録画を終了するには、タスクトレイの停止ボタン(赤く表示される四角形)をクリックします。



6. 録画ファイルに名前を付けて保存します。名前は、「01」「02」のように、録画の順序が分かるような名前にしてください。
7. 録画内容を再生するかどうか確認するダイアログボックスが表示されたら、「いいえ」をクリックします。


録画はポストエディティングするファイルごとに区切る必要はありませんが、数ファイルごとに保存して新しく録画を開始してください。また、休憩する際も、いったん録画ファイルを保存して、再開するときは新しく録画を開始してください。

3. ポストエディティング

➤ ショートカットキー

ポストエディティングで使用していただくキー操作は、標準的な TagEditor のキーシーケンスです(以下の表を参照)。

! セグメントのオープンとクローズは、メニューやツールバーから選択せず、必ず以下のショートカットキーで行ってください。オープンとクローズ以外はマウス操作でもキー操作でもかまいません。

キー操作	動作
Alt + Home	開いて取得する
Alt + End	登録して閉じる  Alt+Shift+End は使用しないでください。
Alt + テンキーの + または Alt + Ctrl + N	登録して閉じ、次のセグメントを開いて取得する

! Oak_Key ユーティリティにより、セグメントを閉じるたびに「\$」記号が自動的にセグメントの末尾に 1 つずつ追加挿入されます。これはデータ解析に必要な記号ですので、絶対に削除しないでください。

! セグメントをクローズするキー操作は、現在開いているセグメント内にカーソルを置いた状態で実行してください。

！ PE 作業中は、必ず Insert(挿入)モードを使用し、Overwrite(上書き)モードにはしないでください。上書きモードになっていると、マクロが正しく動作しないことがあります。

<必ずテストしてください>

前述の「ツールの設定」に従って設定を行った後、これらのキー操作を、まず_Test_kit フォルダに入っている v17219425_ja_jp_17327647.xml ファイルを使用して試してみてください。その際、以下のことを試してください。

- a. 適当に、「あああ」「xxx」など入力したり削除したりしてみる。
- b. すべてのセグメントを一度は開閉する。
- c. 最後までいったら、ランダムにいくつかのセグメントに戻って開閉してみる。
- d. IME ON 全角 / IME ON 半角 / IME OFF のそれぞれの状態でセグメントを開閉してみる。
- e. カーソルを現在のセグメント内のいろんな場所(先頭、末尾、英語内、日本語内など)においてからセグメントを閉じてみる。

➤ テスト時に確認していただきたいこと

1. セグメントを最初に開いたときには末尾に「\$」が 1 つだけ付いている。
2. 閉じるときにもう 1 つ追加されて「\$\$」となる。
3. 同じセグメントを何度か開閉すると、そのたびに「\$」が 1 つずつ追加される。
4. 上記手順の a～e のどの場合でも同様に動作する。

- このように動作しない場合はご質問ください。

➤ ポストエディティングの流れ

1. TagEditor で_Pause_Restart.xml ファイルを開きます。
2. TagEditor でポストエディティング対象の xml ファイルを開きます。

！ PE 中は常に作業対象のファイルと_Pause_Restart ファイルの両方を TagEditor 内で開いておきます。

3. PE 対象の xml ファイルに移動して、ポストエディティング作業を実行します。
 - a. ファイルの先頭にある「Start ファイル ID」セグメントを開いて表示される訳をそのまま受け入れて閉じます。
 - b. ファイル内容を PE します。

- c. 最後に末尾にある「End ファイル ID」セグメントを開いて表示される訳をそのまま受け入れて閉じます。
4. 途中、休憩などで作業を一時停止する場合や複数の日付にまたがって作業を実行する場合など、ひとつのファイル内での作業中に数秒間以上作業を中断する際は、必ず _Pause_Restart ファイルの「Pause」行で Alt+Home を押し、次に Alt+End を押して、セグメントを開閉してください。
5. 作業を再開する時には、必ず _Pause_Restart ファイルの「Restart」行で Alt+Home を押し、次に Alt+End を押して、セグメントを開閉してください。
6. PE が終了したら、[ファイル]メニューの[別名(バイリンガル)で保存]をクリックし、ttx 形式で保存します。

TM には、機械翻訳結果と、翻訳者による翻訳メモリが混在しています。機械翻訳結果の作成者は「MT!」と表示され、翻訳メモリの作成者は「PAL」と表示されます。

機械翻訳結果はすべて 74%一致と表示されますが、この数値は自動的に割り当てられる値であり、実際の翻訳品質を示すものではありません。

🔴💡 日本語文の末尾に付いている「\$」記号は削除しないでください。

☒ TagEditor 内で複数のファイル間を移動するには、Ctrl+F6 が便利です。

< 注意 >

❗ PE 後のファイルはバイリンガル ttx 形式で保存し、そのまま納品してください。xml 形式で保存しないでください。

❗ _Pause_Restart ファイルも常に ttx で上書き保存してください。

🔴💡 クリーンアップは絶対にしないでください。

❗ ポストエディティング作業中は、セグメントを閉じたらすぐに次のセグメントを開くようにし、連続して作業を行ってください。ひとつのファイルの途中で中断するときは、必ず速やかに _Pause_Restart ファイルの「Pause」セグメントを開閉し、再開するときは「Restart」セグメントを開閉してください。

❗ 作業中に以前のセグメントに戻って追加のポストエディティングを行ってもかまいません。

❗ 原文や訳文を読んで確認するだけで、変更を加えない場合でも、必ずセグメントをいったん開いて確認し、確認が終わったら閉じてください。

- ！ 途中で操作を誤ってしまった場合 (Pause / Restart セグメントの開閉し忘れなど) は、影響を受けたと考えられるセグメントの原文を、テキストエディタや MS ワードなどにコピー / ペーストして保存し、他の納品物と一緒にお願いします。
- ！ 一旦作業を終えたファイルを再度開いて編集する場合は、作業の開始時にファイル先頭の「Start ファイル ID」セグメントを開閉し、終了時に末尾の「End ファイル ID」セグメントを開閉してください。

PE 作業終了後、**アンケート.doc** を開いて、回答を記入してください。

III. ポストエディティング ガイドライン

以下は、ポストエディティング時に注意していただきたい点です。これはおおまかな考え方を示すガイドラインであり、個々のケースをすべてカバーするものではありません。

< PE された訳文の品質について >

PE が終了した訳文は、「読者が読んだときにすぐに理解できる」ような文でなければなりません。そのためには、**原文の意味を正しく伝えていることと、日本語の文法に沿っていることが必要とされます。**

しかし、PE にはスピードも求められます。時間をかけて、日本語として美しい文にする必要はありません。そのため、スタイルの洗練などの編集作業は避けてください。

A. 修正が必要な項目

1. 本来翻訳が不要な箇所(コマンドや変数など)が翻訳されている場合 原文に戻す
2. 一般 IT 用語として不適切な語
3. 誤訳 (訳文が原文の意図を正しく伝えていない)
4. 誤訳ではないが語順が不適切で読解が困難な文
5. 意味は通じるがあまりに不自然または不適切な表現
6. 助詞や活用が不適切な文

B. 一貫していることが望ましい項目

1. 操作手順の表現など繰り返し使われる部分
2. マニュアルやヘルプの項目タイトルのスタイル

C. 修正が不要な項目

1. 箇条書きの語尾(体言止めと「です」「ます」が混在してもかまいません)
2. &ProductNameShort;、&ProductNameLong; などには後で製品名が入りますので、このままで結構です。
3. 句読点

以下に、修正が不要な例を具体的に示します。

原文	MT(このままで OK)	このように修正する必要はありません
Other options are:	他のオプションは次のとおりです:	他のオプションは次のとおりです。
Be careful when you add volumes to a disk pool.	ディスクプールにボリュームを追 加するとき注意してください。	ディスクプールにボリュームを追加する 場合は注意してください。

In OpenStorage, NetBackup media servers function as the data movers.	OpenStorage では、NetBackup メディアサーバーはデータムーバーとして機能します。	OpenStorage では、NetBackup メディアサーバーがデータムーバーとして機能します。
Use the array software utilities to delete the volume from the array.	アレイからボリュームを削除するためにアレイソフトウェアユーティリティを使ってください。	アレイソフトウェアユーティリティを使って、ボリュームをアレイから削除します。

D. 製品用語について。

製品用語は、<guimenuitem>タグで囲まれている場合や、[English]形式で記載されている場合があります。また英語のまま残されている場合と日本語に訳されている場合があります。これらの処理については、以下を参照してください。

- ❖ <guimenuitem>[English]</guimenuitem>
機械翻訳の場合: そのままにしておいてください。
TM の場合: 新しい英文の製品用語と一致していることを確認してください。
- ❖ <guimenuitem>日本語または English</guimenuitem>
そのままにしておいてください。不自然な訳が当てられていても編集の必要はありません。[]の挿入も必要ありません。

タグに囲まれていないが製品用語であると思われるもの。またはよくわからない場合。用語は表示されるままにして、日本語 または English のように、で囲んでください。

E. その他の不明点について

今回のポストエディティングのポイントは、訳文が**原文の意味を正しく伝えていることと、日本語の文法に沿っていること**を確認する点です。それ以外の、語尾の統一性やスペースの個数、記号の種類といった問題についてはあまり気にする必要はありませんので、適宜判断して処理していただいて結構です。どうしても判断に困った場合は、訳文内に で囲んでコメントを記してください。

以上、どうぞよろしくお願いいたします。

Oak Project Manual

For the participant post-editors of Oak project

July, 2009

Rev.1.4

Thank you for participating in the post-editing research project.

This document includes three sections:

- I. Project summary
- II. Task instructions
- III. Post-editing guidelines

Please read the entire manual before you start post-editing.

If anything is unclear or does not work as explained in this manual, please contact the researcher before you continue post-editing.

I. Project summary

< Purpose >

The goal of this project is find out what are causing more load on post-editing task, such as the source text characteristics and machine translation errors. In order to do so, we are aiming at measuring the post-editing time on a segment by segment basis, to analyse the characteristics of such segments.

< Handling of data and personal information >

We will record the time and the screen activities to find out the possible problems of the text, and not for judging the ability of each post-editors. The data obtained will only be used solely for the purpose of this research, and any information that may help to identify the participant post-editors will not be publicised.

< About the post-editing material >

The text for post-editing has been extracted from the user manual of a Symantec product.

< Post-editing speed >

You will be asked to post-edit the 32 files of Japanese text, which have been translated from 5,029 word English text. There is no time restriction for the task, although we roughly estimate the post-editing duration as approximately one full day. Please follow the post-editing guidelines at the end of this project manual, and work at a normal speed as you do in your everyday work, keeping the same pace as much as possible. You can have breaks as many times as you like.

< Special procedures >

You are required to exercise a few extra procedures in addition to the standard Trados TagEditor key sequences in order for the precise data collection. Please refer to the next section "Task instructions" and practice the procedure with test files (files in "_Test_kit" folder). If anything does not work as explained, please contact Midori Tatsumi before you begin the post-editing task for the project.

1. Files and folder structure in “Oak kit”

はじめにお読みください_Oak.pdf (This file)

アンケート.doc (Questionnaire file. Please fill in at the end of the project.)

_Test_kit [Files for test and practice]

_Pause_Restart.xml (A file to be used when you pause the post-editing task)

Test_kit.* (TM for Trados 7)

Test_kit65.* (TM for Trados 6.5)

v17219425_ja_jp_17327647.xml (A file for practice)

PE [Files for a take]

*.xml (Files to be post-edited)

_Pause_Restart.xml (A file to be used when you pause the post-editing task)

Settings [Files for various settings]

Oak_Key.exe (To be activated by double-clicking before post-editing)

Ini [Setting files]

TM [Translation memory files]

OAK.* (TM for Trados 7)

OAK65.* (TM for Trados 6.5)

Please do not alternate the folder structure.

Please deliver all the files in the structured folders after finishing the project.

2. Setting up the tools

➤ Settings in TRADOS Workbench

Please start TRADOS Translator's Workbench before you start post-editing, and do the following.

- Specifying the ID
 1. Close the TM if any TM is open.
 2. Select [User ID] from [Settings] menu.
 3. Enter "**PExx**" and click [OK], Where xx is the ID to be assigned to you.
- Opening TM
 1. Select [Open] from [File] menu, choose **OAK.tmw** (**OAK65.tmw** if using Trados 6.5), and click [Open].
- Specifying TM options
 1. Select [Translation Memory Options] from [Options] menu.
 2. In [General] tab, set the [Minimum match value] to 70%.
 3. Select [Merge] for both of the [Updating attribute and text fields] options.
 4. Set the options in [Penalties] tab as follow. Note that [Machine translation penalty] needs to be set to 26%.

Formatting differences penalty %:	1
Attribute and text field differences penalty %:	2
Placeable differences penalty %:	2
Alignment penalty %:	3
Machine translation penalty %:	26
Multiple translations penalty %:	0

Apply placeable penalty also when source tags differ ☐

5. Click [Tag Settings] button in [Tools] tab, click [Open], select Oak_kit\Setting\ini\docbookx.ini and click [Open].
6. Set this file as [Default] and click [OK] twice.

Now you have completed setting Trados Workbench.

When you open the files in TagEditor, please use the ini file you selected here (shown as 'Symbook').

➤ Activation of a macro

A macro called Oak_Key needs to be used in combination with TagEditor. This makes it possible to record thorough time data by only using the standard key sequences of TagEditor.

- a. Double-click **Oak_Key.exe** in 'Settings' folder.
- b. Make sure that Windows task tray shows the [H] icon.



Now you have completed setting the macro.

- ☒ After finishing the post-editing task, please deactivate the macro by right-clicking the [H] icon and selecting [Exit].

➤ Activation of BB FlashBack 2 Express

BB FlashBack is a software programme that records screen activities. We use this programme to record what have been done during the post-editing for detailed analysis of post-editing process.

1. From the [Start] menu, select [Programs], [Blueberry Software], [BB FlashBack 2 Express (Japanese)], [BB FlashBack 2 Express Player].
2. Click the down arrow in Windows task tray and select [Quick record].



3. Click [OK].
4. Recording starts.
5. To stop recording, click the red square in the task tray.



6. Save the record file with names that show the sequence, such as: 01, 02, and so on.

7. If a dialog box appeared and asked if you want to replay the recording, click [No].

Please save the recording for every few post-editing files, and start a new recording.


Also, save the recording before you take a break, and start a new one after the break.

3. Post-editing

➤ Shortcut key sequences

You are required to use standard Trados TagEditor shortcut key sequence. (See the table below.)

! Please make sure you use the following shortcut keys when opening and closing the segments in TagEditor. Please do not select the corresponding options from menus or toolbars. Either mouse or key sequence can be used for operations other than opening and closing the segments.

Key sequence	Action
Alt + Home	Open/Get
Alt + End	Set/Close  Do not use Alt+Shift+End
Alt + '+' in keypad or Alt + Ctrl + N	Set/Close Next Open/Get

! Oak_Key macro adds one '\$' symbol at the end of the segment every time a segment is closed. Please do not delete these symbols as they are necessary for the data analysis.

! Use the key sequence above while the cursor is in the segment that is currently open.

! Make sure you are in the Insert mode, and not in the Overwrite mode, at all time during post-editing. The macro may not work in Overwrite mode.

< Please test >

After completing the settings discussed in the section '2. Setting up the tools', please test above key sequences using the file 'v17219425_ja_jp_17327647.xml' in '_Test_kit' folder. When doing so, please test the following.

- a. Enter and delete some text, such as 'あああ' and 'xxx'.
- b. Open and close all segments at least once.
- c. When finished opening and closing all the segments, return to random segments, and try opening and closing them.
- d. Open and close segments with various settings including IME ON Full-width / IME ON Half-width / IME OFF.
- e. Open a and close it while having the cursor in different positions within the segment, such as the beginning, end, English segment, and Japanese segment.

➤ Please confirm the following:

1. There is only one '\$' when opening a segment for the first time.
2. An '\$' is added when the segment is closed.
3. When a segment is opened and closed more than once, an '\$' is added every time the segment is closed.
4. All of above happens in all conditions listed from a to e above.

- Please contact the researcher if it does not work as explained here.

➤ Post-editing workflows


1. Open the file **_Pause_Restart.xml** in Trados TagEditor.
2. Open an **xml** file to post-edit.

! Keep **_Pause_Restart.xml** always open along with the file(s) to post-edit in TagEditor.

3. Post-edit the xml file.
 - a. Open the first segment '**Start file_ID**', where '*file_ID*' is an ID for each file, and close it without changing the translation.
 - b. Post-edit the contents of the file.
 - c. Open the last segment '**End file_ID**', where '*file_ID*' is an ID for each file, and close it without changing the translation.


4. When you interrupt working more than a few seconds, for having a break, finishing work for the day, and so forth, please open the 'Pause' segment in _Pause_Restart.xml file by pressing Alt+Home, and close it by pressing Alt_End.
5. When you resume working, please open the 'Restart' segment in _Pause_Restart.xml file by pressing Alt+Home, and close it by pressing Alt_End.
6. When you have finished post-editing an xml file, save it in ttx format by clicking [Save Bilingual As] from [File] menu.

The TM includes both machine translation output and the past translation done by human. Machine translation output is marked as 'MT!' and human translation as 'PAL' in Trados. All machine translation outputs are shown as 74% matches, which is automatically assigned by Trados, and does not reflect the quality of translation.

 Please do not delete '\$' symbol at the end of Japanese segments.

☒ You can press Ctrl+F6 to move quickly between multiple files in TagEditor.

< Caution >

- ! Save the post-edited files in the bilingual ttx format, and deliver them as they are. Please do not save them in the monolingual xml format.
- ! Save _Pause_Restart.xml file also in ttx format.
-  Please do not clean up the ttx files.
- ! During the post-editing, be sure to edit the segment in a continuous manner opening the next segment right after closing the previous segment. When you interrupt working in a file, open the 'Pause' segment in _Pause_Restart.xml file by pressing Alt+Home, and close it by pressing Alt_End. When you resume, open and close the 'Restart' segment in the same manner.
- ! You can go back to previously edited segments within the same file any time.
- ! Please open the segments even when you only read and confirm the correct translation without making any changes, and close it after you finished reading.
- ! When you failed to perform above explained procedures, such as opening and closing the 'Pause' and 'Restart' segments when taking a break, please copy and paste the affected segments in a text or MS Word file and deliver it along with other deliverables.
- ! When you reopen a previously edited file for further confirmation or modification, please open and close the 'Start *file_ID*' segment before you start working, and open and close the 'End *file_ID*' segment when you have finished working.

Please fill in the questionnaire file (アンケート.doc) after you have finished post-editing.

III. Post-editing guidelines

Following are the guidelines for post-editing. They only briefly show the basic approach in post-editing, and do not cover every individual case.

< Required quality of post-edited text >

The post-edited text needs to be easily understandable by the readers. In order to achieve that goal, the text needs to **convey the correct meaning of the source text**, and **conform to the Japanese grammar**.

However, speed is another important requirement for post-editing processes. Therefore, it is not necessary to spend time to aesthetically refine the text; please avoid editing for stylistic sophistication.

A. What needs to be fixed:

1. Non-translatable items, such as command and variable names, that have been translated. Please put it back to English.
2. Inappropriately translated general IT terms.
3. Mistranslation (The meaning of the source text has not been conveyed correctly into translation).
4. Word orders that are inappropriate to the level that the sentence has become impossible or difficult to comprehend.
5. Comprehensible but extremely unnatural or inappropriate expressions.
6. Inappropriate postpositions and conjugations.

B. What are preferred to be standardised:

1. Repetitive items, such as procedures and steps.
2. Styles of the section titles in a user manual and help files.

C. What does not need to be fixed:

1. Ending styles of bulleted items. (It is acceptable to have nominal endings and sentence endings mixed together.)
2. Placeholders, such as &ProductNameShort; and &ProductNameLong; as they will be replaced with actual names.
3. Punctuation.

Following is the list of the items that do not have to be fixed.

Source text	Acceptable MT output	Example of unnecessary post-editing
Other options are:	他のオプションは次のとおりです:	他のオプションは次のとおりです。
Be careful when you add volumes to a disk pool.	ディスクプールにボリュームを追加するとき注意してください。	ディスクプールにボリュームを追加する場合は注意してください。
In OpenStorage, NetBackup media servers function as the data movers.	OpenStorage では、NetBackup メディアサーバーはデータムーバーとして機能します。	OpenStorage では、NetBackup メディアサーバーがデータムーバーとして機能します。
Use the array software utilities to delete the volume from the array.	アレイからボリュームを削除するためにアレイソフトウェアユーティリティを使ってください。	アレイソフトウェアユーティリティを使って、ボリュームをアレイから削除します。

D. How to process product specific terms:

The product specific terms are enclosed by <guimenuitem> tags, and in some cases, shown in [English] format. Some product specific terms are left in English, and some are translated into Japanese. Please follow the instructions below for processing these terms.

- ❖ <guimenuitem>[English]</guimenuitem>

For MT output: Please leave it as is.

For TM matches: Please make sure it matches with the new English segment.

- ❖ <guimenuitem>Japanese / English</guimenuitem>

Please leave it as is, even if it is translated into a seemingly inappropriate term.

You do not have to change it to [] format either.

For those terms that are not enclosed by <guimenuitem> tags but appear to be product specific terms, or when you are not sure, please leave the translation as is, and mark the term with ‘ ’, such as Japanese or English .

E. Other

The goal of post-editing in this project is to make the target text **convey the correct meaning of the source text**, and **conform to the Japanese grammar**. For any issues that are beyond these requirements and not covered in this guideline, please handle as you think appropriate. If you have trouble finding a solution, please insert a comment in the target text enclosed in .

Thank you for your participation.

Appendix B PART OF SPEECH LIST

Abbreviation	Full name	Explanation	Japanese	Example (applicable parts are underlined)
Noun	Noun	General nouns, such as 'collection' and 'item'.	一般名詞	<u>コレクション</u> に <u>アイテム</u> を追加します / 定義した方法に応じて
Noun-Sa	Stem for Sa-Verb	Nouns that work as a stem of a Sa-verb.	サ変接続名詞	<u>保持</u> された / <u>定義</u> する / <u>適用</u> される
N-proper	Proper noun	Proper nouns, mostly tags and product specific names in the case of our test corpus.	固有名詞	<guimenuitem> [<u>Custodians</u>] </guimenuitem> <u>タブ</u> をクリック
Pron	Pronoun	Pronouns, such as the Japanese equivalent of 'these' and 'that'.	代名詞	<u>これら</u> の設定 / <u>それ</u> はまた / <u>だれ</u> でも
Noun-adv	Adverbial noun	Nouns that work as an adverb, such as the Japanese equivalent of 'every time' and 'all'.	名詞(副詞的用法)	過去に / 毎回 / すべて
Noun-adjv	Stem for adjective verb	Nouns that work as stems to form adjective verbs, such as the Japanese equivalent of 'necessary' and 'possible'. (An 'adjective verb' is a word that functions as an adjective but conjugates like a verb.)	名詞(形容動詞語幹)	<u>必要</u> です / <u>可能</u> でない / <u>グローバル</u> な
Number	Number	Numbers, such as '1' and '24', and some determiners, such as the Japanese equivalent of 'some'.	数詞	<u>何人</u> かの人々 / <u>24</u> 時間以内 / <u>一度</u> でも
Noun-dep	Dependent noun	Nouns that have meanings by being connected to other words, such as the Japanese equivalent of 'an instance (of an action)', 'when (doing something)', and 'for the purpose of'.	非自立名詞	した <u>こと</u> がある / する <u>とき</u> / その <u>ため</u> の
Noun-suff	Suffixal noun	Nouns that are added to the end of other words to add meanings, such as the Japanese equivalent of '(one) time', '(mark)-ing', and 'within'	接尾語名詞	一度 / マーク <u>付け</u> / コレクション <u>内</u>

Pref	Prefix	Words that are added at the beginning of other words to add meanings, such as the Japanese equivalent of 'every', 'un-', and Japanese specific prefix to add politeness.	接頭詞	<u>各</u> アイテム / <u>未</u> 解決 / <u>ご</u> 用意
Verb	Verb	Verbs, such as the Japanese equivalent of 'capture' and 'use'.	動詞	サンプルを <u>取り込んで</u> / スケジュールを <u>使う</u>
Verb-dep	Dependent verb	Verbs that have meanings by being connected to other words, such as the Japanese equivalent of 'have (someone) to do' and '(accidentally) have done'.	非自立動詞	付いて <u>いる</u> / して <u>もらう</u> / 削除して <u>しまった</u>
Verb-suff	Suffixal verb	Verbs that are added to the end of other words to add meanings. In the case of our test corpus, they are mostly the passive voice element of verbs.	接尾語動詞	表示 <u>される</u> / 受け入れ <u>られた</u> / 適用 <u>されます</u> が
Sa-Verb	S-consonant irregular conjugation verb	Special case verbs that basically mean 'do' and form a 'do+noun' type verb by being added to the end of nouns. Although Sa-verb can be a stand-alone verb, we treated Sa-verbs as function morphs, since an alteration of Sa-verb causes a functional change as opposed to a semantic change.	サ変動詞	保持 <u>された</u> / 保持 <u>します</u> / マーク付け <u>する</u>
Adj	Adjective	Adjectives, such as the Japanese equivalent of 'new', 'fast' and 'not'.	形容詞	<u>新しい</u> 検索を実行 / 最も <u>速い</u> 方法 / 必要が <u>ない</u>
Adj-dep	Dependent adjective	Adjectives that have meanings by being connected to other words, such as the Japanese equivalent of 'easy to...' and 'Ok to...'.	非自立形容詞	わかり <u>やすく</u> 示す / 割り当てても <u>いい</u> です
Adv	Adverb	Adverbs, such as the Japanese equivalent of 'already', 'beforehand' and 'then'.	副詞	<u>すでに</u> マーク付けされて / <u>あらかじめ</u> 受け入れられて / <u>次に</u> [OK] をクリックします
Adv-p	Adverb connected to particles	Adverbs that are connected to particles and the likes, such as the Japanese equivalent of 'whether', 'immediate(ly)', and 'over(ly)'.	助詞類接続副詞	あるか <u>どうか</u> / <u>すぐに</u> / <u>あまり</u> にも
Pren-adj	Prenoun adjectival	Modifiers that modify nouns, do not conjugate, and cannot be a subject. Most cases in the test corpus are equivalent of English demonstrative pronouns, such as 'this' and 'that', or indefinite articles, such as 'any' and 'all', used to modify nouns.	連体詞	<u>この</u> 名前を / <u>その</u> 場合に / <u>どんな</u> ケースにも

Conj	Conjunction	Conjunctions, such as the Japanese equivalent of 'for example', 'also', and 'and'.	接続詞	<u>たとえば</u> 、/ <u>また</u> 個別レベルでも/ <u>メールアドレス</u> <u>および</u> <u>グループの</u>
P-case	Case particle	Particles that are attached to nouns or words that act as nouns, and show the relationship between the preceding and succeeding words, such as indicators of subject, object, and possessives.	格助詞	マーク <u>が</u> どのように/ <u>すべての</u> <u>アイテム</u> / <u>コレクションを</u> 保持します
P-conj	Conjunction particle	Particles that are attached to predicates, and show the conjunctive relationship between words before and after the particle.	接続助詞	設定に応じて、/ <u>していなければ</u> 、/ <u>適用されま</u> <u>すが</u> 、
P-adv	Adverbial particle	Particles that are attached to nouns, words that act as nouns, and other particles, and modify them in the same way as adverbs.	副助詞(係助詞を含む)	どのように機能する <u>か</u> / <u>どんな</u> ケースに <u>でも</u> / <u>そのマークは</u>
Adjv-con	Adjective verb conjugation	Conjugations of adjective verbs.	副詞化助詞	どのように <u>に</u> / <u>個別に</u> / <u>必要に</u> 応じて
Aux-verb	Auxiliary verb	Auxiliary verbs, that are attached to other words and add grammatical meanings, such as affirmative and politeness.	助動詞	保持されたマーク/ <u>どのよう</u> に/ <u>クリックしま</u> <u>す</u> 。
Symb	Symbols	Symbols. In the case of our test corpus, mostly the symbols post-editors have inserted to mark questioned items.	一般記号	<u>_questioned_</u> または <u>_reviewed_</u> など
Punc	Punctuation	Japanese commas and full-stops.	句読点	新しい検索を実行し、/ <u>リンクを作成します</u> 。
Paren	Parenthesis	Parentheses.	括弧	<guimenuitem> <u>[Custodians]</u> </guimenuitem> <u>タブをクリック</u>

Appendix C POST-ESTIMATION TESTS

This appendix tests the validity of the data and the model used in the multiple regression analysis in Chapter 7 (section 7.6) from various view points.

1 Outliers, Leverage, and Influence

In a linear regression analysis, if an observation has a dependent-variable value, PE speed in the case of present study, that is unusually large or small compared to its predicted value, it is considered as an *outlier*. In order to detect outliers, we calculated the *Studentized residual*⁸¹ of regression analysis results. If an observation's value of Studentized residual exceeds 2, it is regarded as an outlier. In the present study, we performed the test on the base model (Model II) discussed in section 7.6, and found 105 such cases out of all 2,391 observations. These observations have been further examined in the explanatory phase (Chapter 8) to find out what caused exceptionally fast or slow PE speed.

In a linear regression analysis, if an observation has an extremely large or small value, it is considered as *leverage*. Leverage points may influence the estimate of regression coefficients, and consequently, distort the results of the regression analysis, thus need to be either eliminated or controlled. The leverage can be diagnosed by statistics software, and in the present study, no such instance was detected from our regression results, thus we consider there is no leverage issue in our analysis.

2 Heteroscedasticity

One of the assumptions for the linear regression analysis is that the residuals are distributed homogeneously, that is, homoscedastic. If the distribution has a certain pattern, it is said to be *heteroscedastic*, and indicates that the model is not well fitted. In order to check if there is a heteroscedasticity problem in our base model, we first employed a graphical method, which is shown in Figure C.1. We see that the

⁸¹ Residual is the difference between the predicted value from the linear regression analysis and the observed value.

distribution of the data points is generally symmetrical around the 0 line, but getting sparse towards left and right ends, which is an indication of heteroscedasticity.

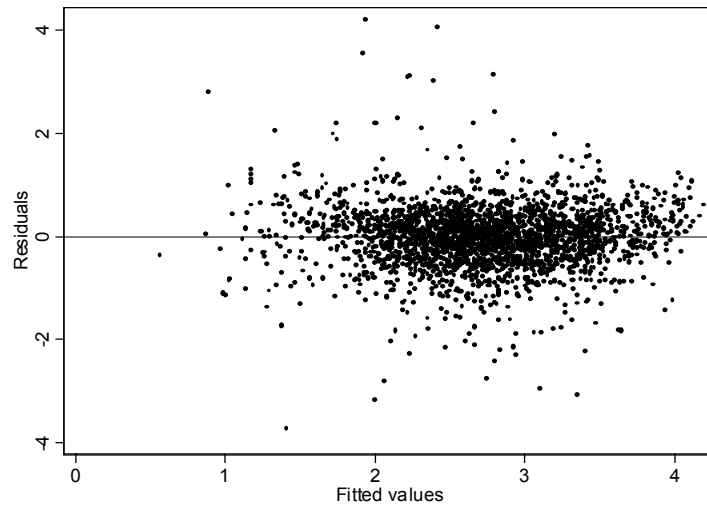


Figure C.1 Heteroscedasticity test

We also ran Breusch-Pagan test, which tests the null hypothesis that the variance of the residuals is homogeneous, and obtained the probability value $p < 0.001$; the evidence was against homoscedasticity. Together with the graphical test we concluded there was a heteroscedasticity problem and decided to take measures to address the issue by means of White's robust standard errors, which adjust the standard errors taking into account the heteroscedasticity problem. The regression analysis results shown in section 7.6 is the outcome of employing the robust option.

3 Multicollinearity

When two or more independent variables have perfect or strong correlation, it is called multicollinearity. Although having independent variables that do not correlate with each other can explain the variance of dependent variables better, multicollinearity itself does not violate the assumption of multiple linear regression analysis, and whether multicollinearity is a problem is open to debate (Wooldridge 2006). Nonetheless, we employed a Variance Inflation Factor (vif) test to examine if multicollinearity existed in our base model, especially since we detected a high correlation between ST length and Systran complexity index (section 7.4.6). As a result, although ST length and Systran complexity index had rather high correlation, it did not exceed the cut-off point set for

the vif measure. There was no other instance that showed high correlation except for the ST length and the square of ST length, which are multicollinear by design.

4 Normality of Errors

Although multiple linear regression does not require normal distributions of either independent variables or residuals (errors), testing the normality of errors help us to determine how well the estimated model fits to the actual, observed data. There are two types of tests: 1) a probability-probability plot (P-P plot), which is sensitive to non-normality in the middle range of data, and 2) quantile-quantile plot (Q-Q plot), which is sensitive to non-normality near the tails. Figure C.2 shows the results respectively.

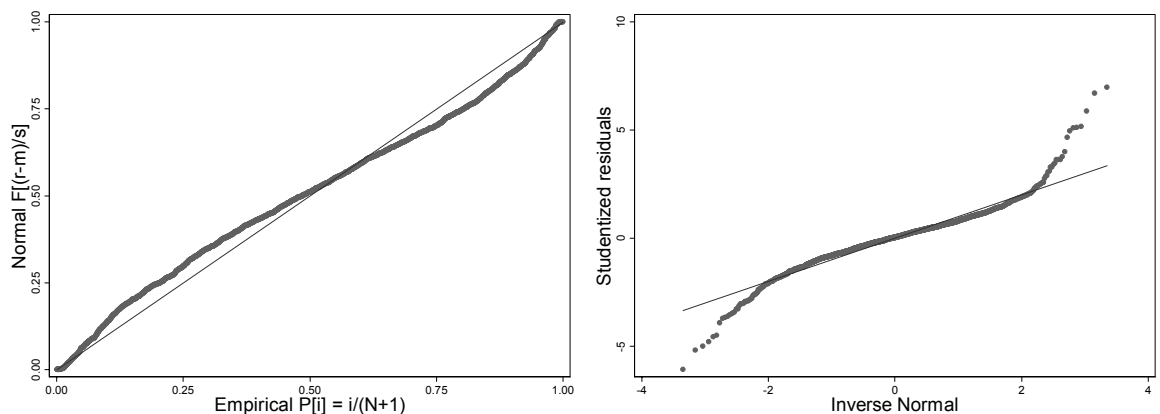


Figure C.2 P-P plot (left) and Q-Q plot (right)

They both show non-normalities to some extent, but neither of them are regarded as major deviation, thus we can accept that the residuals are close to a normal distribution.

5 Model Specification Error

In order to examine if our base model is missing any significant independent variables, we performed the Omitted Variable test. The result indicated that further variables would improve the model.⁸² We speculated that some of the variances are post-editor dependent, thus we tested the model separately on each post-editor. The results showed

⁸² The result of this test is represented as a probability value of the null hypothesis 'The model has no omitted values.' Our result was $p < 0.001$, thus indicated that we could not reject this hypothesis.

that the model did not prove to have model specification errors for 6 out of 9 post-editors, thus being aware there may be more independent variables that can further explain the dependent variable, we believe our base model still gives insight into various effects on PE speed in general.

6 Summary

In this appendix, we have performed various post-estimation tests. In order to address the detected heteroscedasticity issue, we employed the robust option for our regression analysis in section 7.6. We also spotted some outliers, which have been further examined in Chapter 8, however, those data points are not regarded as influential as a result of leverage diagnostics. Our model did not show either multicollinearity problem or severe non-normality of errors. The model specification test showed that there may be other independent variables that have significant effect on PE speed, which left further tasks in PE research.

Appendix D QUESTIONNAIRE

This appendix includes the actual questionnaire used for the project (in Japanese), and its English translation (translated by the author of this thesis).

調査にご協力いただいた方へのアンケート

Oak プロジェクト - 2009 年 7 月

このアンケート用紙は、印刷せず、MS Word で開いて入力してください。

選択回答部分はクリックすると ☒ に変わります。記入回答部分は、括弧内のグレーの部分をクリックしてご記入ください。表示されている括弧は小さいですが、記入とともに拡大します。なるべく詳しくお答えくださいますよう、お願いいたします。

Trados に設定した ID をご記入ください (例:PE01): []

翻訳経験について

1. コンピュータ ソフトウェア マニュアルの翻訳経験はありますか？

☐ はい ☐ いいえ

2. 「はい」とお答えになった場合は、経験年数を選んでください。

☐ 1 年未満 ☐ 1～3 年 ☐ 3～5 年 ☐ 5 年以上

3. コンピュータ ソフトウェア マニュアル以外のテクニカル文書の翻訳経験はありますか？

☐ はい ☐ いいえ

4. 「はい」とお答えになった場合は、経験年数を選んでください。

☐ 1 年未満 ☐ 1～3 年 ☐ 3～5 年 ☐ 5 年以上

5. コンピュータ ソフトウェア マニュアルやテクニカル文書以外の翻訳経験はありますか？

☐ はい ☐ いいえ

6. 「はい」とお答えになった場合は、経験年数を選んでください。

☐ 1 年未満 ☐ 1～3 年 ☐ 3～5 年 ☐ 5 年以上

ポストエディティング経験について

1. コンピュータ ソフトウェア マニュアルの機械翻訳出力のポストエディティング経験はありますか？
☐ はい ☐ いいえ
2. 「はい」とお答えになった場合は、経験年数を選んでください。
☐ 1年未満 ☐ 1～3年 ☐ 3～5年 ☐ 5年以上
3. コンピュータ ソフトウェア マニュアル以外のテクニカル文書の機械翻訳出力のポストエディティング経験はありますか？
☐ はい ☐ いいえ
4. 「はい」とお答えになった場合は、経験年数を選んでください。
☐ 1年未満 ☐ 1～3年 ☐ 3～5年 ☐ 5年以上
5. コンピュータ ソフトウェア マニュアルやテクニカル文書以外の機械翻訳出力のポストエディティング経験はありますか？
☐ はい ☐ いいえ
6. 「はい」とお答えになった場合は、経験年数を選んでください。
☐ 1年未満 ☐ 1～3年 ☐ 3～5年 ☐ 5年以上
7. 機械翻訳出力のポストエディティングのトレーニングを受けたことはありますか？
☐ はい ☐ いいえ
8. 「はい」とお答えになった場合は、トレーニングのタイプと期間を記入してください。(例:「社内トレーニングで2日間」、「翻訳スクールで2週間」など)
[]
9. 7で「はい」とお答えになった場合は、トレーニングで特に有用であった点と有用でなかった点を記入してください。
有用だった点: []
有用でなかった点: []

10. 7 への返答に関係なくお答えください。どのようなポストエディティングトレーニングがあれば有用であるとお考えになりますか？

[]

ツールとフォーマットについて

1. TRADOS TagEditor を翻訳ツールとして使用されたことはありますか？

☐ はい ☐ いいえ

2. 「はい」とお答えになった場合は、経験年数を選んでください。

☐ 1 年未満 ☐ 1～3 年 ☐ 3～5 年 ☐ 5 年以上

3. TRADOS TagEditor をポストエディティングツールとして使用されたことはありますか？

☐ はい ☐ いいえ

4. 「はい」とお答えになった場合は、経験年数を選んでください。

☐ 1 年未満 ☐ 1～3 年 ☐ 3～5 年 ☐ 5 年以上

5. この調査で使用したツール (TagEditor + マクロ + BB Flashback) は問題なく使用できましたか？

☐ 問題なかった ☐ やや問題があった ☐ 非常に問題があった

6. 「問題なかった」以外を選択された場合は、その理由を記入してください。

[]

7. TagEditor で xml 形式のファイルを使用された経験はありますか？

☐ はい ☐ いいえ

8. 「はい」とお答えになった場合は、経験年数を選んでください。

☐ 1 年未満 ☐ 1～3 年 ☐ 3～5 年 ☐ 5 年以上

「ポストエディティングガイドライン」について

1. この調査に使用したポストエディティングガイドラインの指示内容の中で、特に従うのが難しいと思われた点や、ポストエディティングの効率低下の原因になったと思われた点がありますか？
☐ はい ☐ いいえ
2. 「はい」とお答えになった場合は、問題点を具体的に説明してください。
[]
3. この調査で使用したガイドラインに限らず、一般的に、ポストエディティングガイドラインの内容で、特に従うのが難しいと思われる点や、ポストエディティングの効率低下の原因になっていると思われる点がありますか？
☐ はい ☐ いいえ
4. 「はい」とお答えになった場合は、問題点を具体的に説明してください。
[]

ポストエディティングの手法について

1. ポストエディティング作業はどのような流れで行いますか？
例：
 - 原文を読む → 訳文を読む → 訳文を編集する
 - 訳文を読む → 訳文を編集する → 原文と照合する
 - 原文を読む → 原文を頭の中で翻訳する → 訳文を編集する、など

できるだけ詳しく記述してください。

[]
2. 上でお答えになった流れは、原文の種類や訳文の品質などの条件によって変化しますか？
☐ はい ☐ いいえ
3. 「はい」とお答えになった場合は、具体的に説明してください。
[]

4. Trados Translator's Workbench の左側に表示される情報フィールドのうち、普段参考にされるフィールドをすべて選択してください。

- ☐ 作成日
- ☐ 作成者
- ☐ 更新日
- ☐ 更新者
- ☐ 最終使用日
- ☐ 使用カウンタ

機械翻訳と TM について

1. 機械翻訳結果と TM マッチセグメントは、どちらが編集しやすいと思われますか？当てはまるものをすべて選択してください。

- ☐ 一般に機械翻訳結果
- ☐ 一般に TM マッチセグメント
- ☐ どちらともいえない
- ☐ 機械翻訳の品質によって異なる
- ☐ TM のマッチ率によって異なる

その他

1. このプロジェクトに関するご意見や感想などありましたらご記入ください。

[]

アンケートは以上です。ご協力ありがとうございました。

Questionnaire for participant post-editors (Translation)

Oak project - July, 2009

Please open and fill in this questionnaire in MS Word; please do not use the printout for your answers.

For alternatives, please click the box to change it to ☒. For free-answer questions, please click the grey box within the brackets and start typing. The space in the brackets expands as you type. Please answer in as much detail as possible.

Please enter the ID specified in Trados (Ex. PE01) : []

About your experience in translation

1. Do you have experience in translating computer software documentation?
☐ Yes ☐ No

2. If you have answered 'Yes', please select the length of experience in translating software documentation.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years

3. Do you have experience in translating technical documents other than computer software documentation?
☐ Yes ☐ No

4. If you have answered 'Yes', please select the length of experience in translating technical documents other than computer software documentation.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years

5. Do you have experience in translating anything other than computer software or technical documents?
☐ Yes ☐ No

6. If you have answered 'Yes', please select the length of experience in translating anything other than computer software or technical documents.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years

About your experience in post-editing

1. Do you have experience in post-editing computer software documentation?
☐ Yes ☐ No
2. If you have answered 'Yes', please select the length of experience in post-editing software documentation.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years
3. Do you have experience in post-editing technical documents other than computer software documentation?
☐ Yes ☐ No
4. If you have answered 'Yes', please select the length of experience in post-editing technical documents other than computer software documentation.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years
5. Do you have experience in post-editing anything other than computer software or technical documents?
☐ Yes ☐ No
6. If you have answered 'Yes', please select the length of experience in post-editing anything other than computer software or technical documents.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years
7. Have you received training in post-editing machine translation output?
☐ Yes ☐ No
8. If you have answered 'Yes', please specify the type and the duration. (Ex. 'A two-day in-house training session', 'A two-week course given by a translation training school.')
[]
9. If you have answered 'Yes' to question 7, please specify what was and was not useful.
What was useful: []
What was not useful: []

10. Please answer regardless of your answer to question 7. What kind of post-editing training do you think would be useful?
- []

About tools and format

1. Have you used TRADOS TagEditor as a translation tool?
☐ Yes ☐ No
2. If you have answered 'Yes', please select the length of experience in using TRADOS TagEditor as a translation tool.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years
3. Have you used TRADOS TagEditor as a post-editing tool?
☐ Yes ☐ No
4. If you have answered 'Yes', please select the length of experience in using TRADOS TagEditor as a post-editing tool.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years
5. Have there been any problems with tools (TagEditor + Macro + BB FlashBack) used for this study?
☐ No ☐ Yes, to some extent ☐ Yes, very much
6. If you have chosen other than 'No', please specify the reason.
[]
7. Have you used xml files with TagEditor?
☐ Yes ☐ No
8. If you have answered 'Yes', please select the length of experience in using xml files with TagEditor.
☐ Less than 1 year ☐ 1 – 3 year(s) ☐ 3 – 5 years ☐ More than 5 years

About post-editing guidelines

1. Are there any instructions in the post-editing guidelines used in this study that have been difficult to follow or that have caused inefficiency in post-editing?
☐ Yes ☐ No
2. If you have answered 'Yes', please explain what the problems were.
[]
3. In general, not limited to the guidelines used for this study, are there any instructions in post-editing guidelines that are difficult to follow or that cause inefficiency in post-editing?
☐ Yes ☐ No
4. If you have answered 'Yes', please explain what the problems are.
[]

About your post-editing method

1. How do you proceed with post-editing tasks in general? For example:
 - Read source text -> Read translation -> Edit translation
 - Read translation -> Edit translation -> Check against source text
 - Read source text -> Translate source text in the head -> Edit translation

Please explain in as detailed a manner

[]

2. Do you change the method you outline above depending on the types of source text and/or quality of translation?
☐ Yes ☐ No
3. If you have answered 'Yes', please explain in detail.
[]
4. Which information fields displayed in Trados Translator's Workbench do you usually refer to? Please select all applicable items.
☐ Created on:

- ☐ Created by:
- ☐ Changed on:
- ☐ Changed by:
- ☐ Last used:
- ☐ Usage:

About machine translation output and TM

1. Which do you think is easier to edit, machine translation output or TM match segment?

Please select all applicable items.

- ☐ Generally machine translation output
- ☐ Generally TM match segment
- ☐ Difficult to tell
- ☐ It depends on the quality of machine translation output
- ☐ It depends on the match ratio of TM

Other

1. Please provide any opinions and thoughts you might have about this project in general.

[]

This is the end of questionnaire. Thank you for your cooperation.

Appendix E DCU RESEARCH ETHICS COMMITTEE DOCUMENTS

This appendix includes the plain language statement and the informed consent sheet used for the project.

I. Introduction to the Research Study

This is a research project to understand the nature of professional post-editing work and the possible causes that make post-editing an effort intensive task. The main method of this research is compare the types and the amount of changes made on the text and the time taken to make such changes. This project is being carried out by Midori TATSUMI as her PhD research project. Her contact details are:

*School of Applied Language and Intercultural Studies
Dublin City University*

*Telephone: +353 (0)86 409 1645
Email: midori.tatsumi2@mail.dcu.ie*

II. Details of what involvement in the Research Study will require

The research will involve the following:

- 1. Post-editing the machine-translated Japanese document according to the provided guidelines and procedure specifications.*
- 2. Completing a questionnaire about your experience and opinions relevant to post-editing.*

III. Potential risks to participants from involvement in the Research Study (if greater than that encountered in everyday life)

There are no risks involved in participating in this study.

IV. Benefits (direct or indirect) to participants from involvement in the Research Study

The indirect benefits of your participation in this study are that:

- Helping to understand the nature of post-editing process*
- Helping to identify the ways to reduce post-editing effort in the future*

V. Advice as to arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations

Your anonymity will be protected at all times. You will be given an identifier such as "Post-editor 1" and no mention will ever be made of your real identity in the final report. The data collected will be used only by Midori TATSUMI and will not be given to anybody else.

VI. Advice as to whether or not data is to be destroyed after a minimum period

The data will be stored in a secure location only at DCU. The data will be destroyed within five years of its acquisition.

VII. Statement that involvement in the Research Study is voluntary

You will be paid normal working rates for participation in this study as it is important that your post-editing work resembles what you would do in a normal working environment. However, given that this is for a research project, your participation is voluntary, and there will be no penalty for withdrawing from the research study.

**If participants have concerns about this study and wish to contact an independent person,
please contact:**

The Secretary, Dublin City University Research Ethics Committee, c/o Office of the Vice-President for Research, Dublin City University, Dublin 9. Tel 01-7008000

Appendix F RESEARCH CONSENT FORM

This appendix includes the consent form distributed to and signed by the participants.

I. Research Study Title

Comparison and Analysis of Textual Difference and Temporal Effort of Japanese Post-Editing

*Midori TATSUMI, School of Applied Language and Intercultural Studies,
Dublin City University*

II. Clarification of the purpose of the research

The two main purposes of this PhD research project are:

- (1) To have an insight into the nature of professional post-editing work*
- (2) To find out the causes that are making post-editing an effort intensive task*

III. Confirmation of particular requirements as highlighted in the Plain Language Statement

You will be required to post-edit Japanese document that have been partly machine-translated and partly extracted from human-translated translation memory. Approximate amount of text is 5,700 words in source English word count. Post-editing needs to be performed according to the provided guidelines and procedure specifications. You will also be asked to complete a questionnaire about your experience and opinions related to post-editing.

Participant – please complete the following (Circle Yes or No for each question)

<i>Have you read the Plain Language Statement?</i>	Yes/No
<i>Do you understand the information provided?</i>	Yes/No
<i>Have you had an opportunity to ask questions and discuss this study?</i>	Yes/No
<i>Have you received satisfactory answers to all your questions?</i>	Yes/No

IV. Confirmation that involvement in the Research Study is voluntary

You will be paid normal working rates for participation in this study as it is important that your post-editing work resembles what you would do in a normal working environment. However, given that this is for a research project, your participation is voluntary, and there will be no penalty for withdrawing from the research study.

V. Advice as to arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations

Your anonymity will be protected at all times. You will be given an identifier such as “Post-editor 1” and no mention will ever be made of your real identity in the final report. The data collated will be used only by Midori TATSUMI and will not be given to anybody else.

VII. Signature:

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature:

Name in Block Capitals:

Witness:

Date:
