



PRODUCTIVITY AND QUALITY IN THE POST-EDITING OF OUTPUTS FROM TRANSLATION MEMORIES AND MACHINE TRANSLATION

Ana Guerberof Arenas

Dipòsit Legal: T. 1292-2012

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

Ana Guerberof Arenas

**Productivity and quality in the post-editing of outputs from
translation memories and machine translation**

Doctoral Thesis



UNIVERSITAT ROVIRA I VIRGILI

Ana Guerberof Arenas

**Productivity and quality in the post-editing of outputs from translation
memories and machine translation**

Doctoral Thesis

Supervised by

Dr. Anthony Pym, Universitat Rovira i Virgili

Dr. Sharon O'Brien, Dublin City University

PhD program in Translation and Intercultural Studies

Universitat Rovira i Virgili, Spain



UNIVERSITAT ROVIRA I VIRGILI

Tarragona 2012

Abstract

Machine-translated segments are nowadays included as fuzzy matches within the translation memory systems in the localization workflow. This study presents results on the correlation between these two types of segments in terms of productivity, final quality and experience. In order to test these variables, we setup an experiment with a group of twenty four professional translators using an on-line post-editing tool and a customized Moses statistical-base machine translation engine with a BLEU score of 0.6. The translators were asked to translate from English to Spanish, working on no-match, machine-translated and translation memory segments from the 85-94 percent value using a post-editing tool, without actually knowing the origin of each segment, and to complete an on-line questionnaire after the task. These translations were analyzed through a TER score calculation and they were reviewed by three professional translators to assess the resulting quality of the assignment. The findings suggest that translators have higher productivity and quality when using machine-translated output than when translating on their own, and that this productivity and quality is not significantly different from the values obtained when processing fuzzy matches from translation memories in the range 85-94 percent. Furthermore, translators' experience and training seems to have an impact on the productivity and on the quality produced by translators.

Keywords

Translation memory, machine translation, post-editing, review, productivity, quality, errors, editing, professional translators, reviewers, experience, fuzzy match, processing speed, localization, TER, Levenshtein distance, Olivier and Hand index



Professor Anthony Pym
URV. Av. Catalunya 35
43002 Tarragona, Spain
anthony.pym@urv.cat

June 30, 2012

I hereby certify that the present study *Productivity and quality in the post-editing of outputs from translation memories and machine translation*, presented by Ana Guerberof Arenas for the award of the degree of Doctor, has been carried out under the supervision of myself at the Department of English and German Studies of the Rovira i Virgili University and Dr. Sharon O'Brien at the Language Department of Dublin City University, and that it fulfills all the requirements for the mention "Doctor Europeus".

Professor Anthony Pym
Intercultural Studies Group
Universitat Rovira i Virgili
Tarragona, Spain

President
European Society for
Translation Studies

Dublin City University
Ollscoil Chathair Bhaile Átha Cliath



June 30th 2012

I hereby certify that the present study *Productivity and quality in the post-editing of outputs from translation memories and machine translation*, presented by Ana Guerberof for the award of the degree of Doctor, has been carried out under my supervision at Dublin City University, in collaboration with Prof. Anthony Pym at the URV Tarragona, and that it fulfills all the requirements for the mention “Doctor Europeus”.

A handwritten signature in black ink, appearing to read 'S. O'Brien', is written in a cursive style.

Dr. Sharon O'Brien

School of Applied Language and Intercultural Studies

Dublin City University

Acknowledgments

I would like to thank Anthony Pym for his feedback and support, and for directing this doctoral program in Spain, and Sharon O'Brien for her intelligent, clear and timely comments, as well as her encouragement during this thesis and my stay in DCU. Dr. O'Brien has been a source of inspiration for everyone researching in this area. Her work has certainly inspired this research. Thank you both!

I have been blessed to receive the support of so many people! I would like to thank:

Heidi Depraetere and Joeri Van de Walle in CrossLang for their incredible and generous help in training the engine and preparing the corpus, I would be eternally grateful; Ester Boixadera and Anna Espinal in the SEA for their intelligent and professional statistical work, as well as their patience with my questions; Pilar Sánchez Gijón and Olga Torres Hostench for helping with all the UAB resources; Dublin City University for hosting me wonderfully for three months, it was an honour to be there; The CNGL and the EAMT for funding this project, to Fiona Maguire for her kindness and efficiency, and to Andy Way for his interest in the project; Teresa Ortiz, Alba Guix, Charles Lynch and Luigi Riboldi in Logoscript/HiSoft for their flexibility, support and for believing!; MicroStrategy, and especially Laura Porcelli, for providing the corpus; Sergi Martínez for creating Slicer, for Word, and for his knowledge on everything!; Joss Moorkens for his unselfish support on the qualitative section and for introducing me to a wonderful group of PhD students in DCU!; the translation and reviewing team for their contribution, this research would have definitely not been possible without them; Sandra Cascallana and Albert Recolons for being so generous and loving during my trip to Dublin; Marian Bakaikoa for teaching me how to live better, still learning; and my family and friends all over the globe, as well as my *furry* friends for making everything in my life worthwhile!

Table of contents

PART I: Introduction.....	1
Research questions	3
Thesis structure.....	3
Chapter 1: Overview of TM and MT technologies.....	5
1.1. Machine translation and translation memory systems.....	5
1.1.1. Machine translation	5
1.1.2. Translation memory systems	7
Chapter 2: Literature review	9
2.1. Post-editing.....	9
2.1.1. Early studies in post-editing	9
2.1.2. Post-editing and cognitive effort	13
2.1.3. Post-editing and quality	14
2.1.4. Post-editing and experience.....	18
2.1.5. Post-editing and automatic metric scores	20
2.1.6. Post-editing and confidence scores.....	22
2.1.7. Post-editing in a commercial setting	25
2.2. Translation memories: Productivity, quality and processes	32
Chapter 3: Methodological considerations	43
3.1. Pilot Project	43
3.2. Hypotheses	44
3.2.1. Hypothesis 1: Productivity	44
3.2.1.1. Sub-hypothesis	45
3.2.2. Hypothesis: Quality	45
3.2.2.1. Sub-hypothesis	45
3.2.3. Hypothesis 3: Experience	46
3.2.3.1. Sub-hypothesis	46
3.3. Variables and operationalization	46
3.4. Methodology.....	47
3.4.1. Mixed methods approach.....	47
3.4.2. Overview of experimental project	48

3.4.3. Data for quantitative analysis	48
3.4.3.1. Sample	49
3.4.3.1.1. Criteria for selecting translators	49
3.4.3.1.2. Criteria for selecting reviewers.....	49
3.4.3.1.3. Selection process	50
3.4.3.2. Corpus.....	50
3.4.3.3. Statistical Machine Translation (SMT) engine.....	51
3.4.3.3.1. Training the engine	51
3.4.3.4. BLEU score and human evaluation	52
3.4.3.5. Dataset	54
3.4.3.5.1. Selecting the fuzzy matches	56
3.4.3.6. Post-editing tool.....	58
3.4.3.7. Assignment instructions for translators	60
3.4.3.8. Assignment instructions for reviewers:	60
3.4.3.9. The LISA QA Model.....	62
3.4.3.10. Translation Edit Rate (TER).....	62
3.4.3.11. Glossary	64
3.4.4. Data for qualitative analysis	64
3.4.4.1. Translators' questionnaire	64
3.4.4.2. Reviewers' questionnaire	65
3.4.4.3. Debriefings with translators and reviewers	66
3.4.5. Validity and generalizability of the research	68
3.4.6. Threats to validity	70
3.4.6.1. Languages used	71
3.4.6.2. Language variant	71
3.4.6.3. Selection of post-editing tool.....	71
3.4.6.4. Revision by a third party	72
3.4.6.5. Statistical engine.....	72
3.4.6.6. Sequence of segments.....	72
3.4.6.7. Selection of translation memory system.....	73
3.4.7. Testing the methodology	73
3.5. Project development	74
3.5.1. Translation phase	74
3.5.2. Schedule	76
3.5.3. Reviewing phase.....	77

PART II: Quantitative results	81
II.1 Statistical analysis	81
II.1.1 Productivity	83
II.1.2 Quality	84
II.1.3 Experience	85
Chapter 4: Productivity results	87
4.1. Processing time and processing speed	87
4.2. Processing speed by Match category	87
4.3. Productivity gain	93
4.3.1. Estimated words per day	97
4.4. Speed groups and processing speeds	99
4.4.1. Standardized Words per Minute per translator	99
4.5. Speed and number of edits: TER	105
4.5.1. Correlation between TER, Levenshtein, and Olivier and Hand	106
4.5.2. TER indicator: segments edited	107
4.5.3. TER score: edits per segment	110
4.6. Conclusions on productivity	123
Chapter 5: Quality results	125
5.1. The corpus and the type of changes required	126
5.2. The review	129
5.2.1. Revision time	129
5.2.2. Number of errors in review	132
5.2.3. Comparing reviewers	136
5.2.4. Global error database vs. Segment error database	142
5.3. Errors at segment level	144
5.4. Error count	147
5.5. Error classification	152
5.6. Errors vs. processing speed	156
5.7. Overcorrections	162
5.8. Conclusions on quality	165
Chapter 6: Translators' experience results	167
6.1. "About your experience"	167
6.2. Grouping translators according to their experience	173

6.3. Clusters	175
6.3.1. Features of Cluster 1	175
6.3.2. Features of Cluster 2	176
6.3.3. Features of Cluster 3	177
6.3.4. Features of Cluster 4	177
6.4. Experience vs. processing speed	178
6.4.1. Experience vs. processing speed: Fuzzy match	178
6.4.2. Experience vs. processing speed: MT match	180
6.4.3. Experience vs. processing speed: No match	182
6.5. Experience vs. number of errors	186
6.5.1. Experience vs. number of errors: Fuzzy matches	187
6.5.2. Experience vs. number of errors: MT matches	189
6.5.3. Experience vs. number of errors: No matches	191
6.6. Conclusions on the translators' experience	196
PART III: Qualitative results	199
Chapter 7: Questionnaire results	199
7.1. Translators' opinions	199
7.2. Reviewers' opinions	207
7.2.1. Review methodology	207
7.2.2. Pricing	208
7.2.3. Opinions of MT	208
7.3. Translators' opinions of the assignment	209
7.4. Reviewers' opinions of the assignment	213
7.4.1. Translation quality	213
7.4.2. Difficulties in the text	214
7.4.3. The LISA form	215
7.4.4. Assignment review method	215
Chapter 8: Debriefings	217
8.1. Translators' debriefings	217
8.1.1. Assignment	219
8.1.1.1. Instructions	219
8.1.1.2. Glossary	221
8.1.1.3. Questionnaire	222

8.1.1.4. The tool used in the assignment	223
8.1.1.5. Segments.....	224
8.1.2. Feelings.....	225
8.1.3. Machine Translation	227
8.2. Reviewers' debriefings	231
8.2.1. Assignment	232
8.2.1.1. Instructions and methodology	232
8.2.1.2. Glossary	232
8.2.1.3. Questionnaire.....	233
8.2.1.4. Translations	233
8.2.2. Feelings.....	235
8.2.3. Review process	236
8.3. Conclusion on qualitative results.....	237
PART IV: Conclusions.....	241
Chapter 9: Final conclusions	241
9.1. Conclusions of combined results.....	241
9.2. Further research	248
References.....	251
Appendix A.....	265
Glossary	265
Estimation and estimate.....	265
Explanatory and response variables	265
Confidence intervals	265
Continuous and discrete variables	265
Kappa coefficient.....	266
Kruskal-Wallis analysis of variance	266
Fuzzy match.....	266
LISA and the LISA QA Model.....	267
Fuzzy match and MT post-editing pricing	267
Overcorrection.....	269
Poisson distribution	269
Post-editing.....	269
Preferential changes.....	269

Processing speed.....	270
Processing time.....	270
Productivity gain	270
Reviewer.....	270
Translator and post-editor.....	271
Translation Edit Rate (TER).....	271
Variable of interest	271
Appendix B.....	272
Appendix C.....	272
Appendix D.....	272
Appendix E.....	272
Appendix F	272
Appendix G	273
Appendix H	273
Appendix I.....	273
Appendix J	273

List of tables

Table 1: Translation memory components	52
Table 2: Human evaluation criteria	54
Table 3: Analysis of online help for project 9.1	55
Table 4: Analysis of software strings for project 9.1	55
Table 5: Assignment dates and locations	77
Table 6: Descriptive statistics for global processing speed.....	89
Table 7: Estimated global words per minute.....	90
Table 8: Descriptive statistics for Words per minute per translator	92
Table 9 Estimated WPM and productivity gain per individual translator.....	95
Table 10: Estimated words per day	99
Table 11: Standardized WPM with respect to No match per group.....	103
Table 12: Estimated standardized Words per minute.....	104
Table 13: SWPM and speed groups	104

Table 14: SWPM and speed groups & Match categories	105
Table 15: TER indicator per translator	109
Table 16: Descriptive values for TER	111
Table 17: Estimated mean values for TER	111
Table 18: TER descriptive data per segment	114
Table 19: Global TER according to Speed group.....	121
Table 20: Descriptive analysis of Review time	130
Table 21: Descriptive analysis of Review time in WPM	131
Table 22: Mean of total words reviewed in 8 hours	132
Table 23: Percentage of error indicator	133
Table 24: Total number of errors.....	133
Table 25: Descriptive data on errors per reviewer	134
Table 26: Kappa statistical values according to Match category and Reviewer	137
Table 27: Global error database vs. Segment error database	143
Table 28: Differences between databases according to error typology.....	144
Table 29: Error indicator per category	145
Table 30: Odds of “making a mistake”	146
Table 31: Estimated Odds ratio per category	146
Table 32: Error indicator per reviewer	147
Table 33: Number of errors per Match category and translator	147
Table 34: Descriptive analysis of total errors	149
Table 35: Estimated mean of errors in original text	150
Table 36: Number and percentage of errors per type of error	152
Table 37: Reviewer 1 number and percent of errors per type of error	154
Table 38: Reviewer 2 number and % of errors per type of error.....	155
Table 39: Reviewer 3 number and percent of errors per type of error	155
Table 40: Translators’ processing speed versus number of errors	157
Table 41: Errors vs. Speed groups in Fuzzy match	159
Table 42: Errors vs. Speed groups in MT match.....	159
Table 43: Errors vs. Speed groups in No match	160
Table 44: Translators according to three Speed groups	161
Table 45: Overcorrections and Speed groups in Fuzzy match	162
Table 46: Overcorrections and Speed group in MT match	163

Table 47: Overcorrections vs. global errors	164
Table 48: Experience in the localization industry	168
Table 49: Experience in translation memory tools.....	168
Table 50: Experience in business intelligence translation.....	169
Table 51: Translation work for MicroStrategy.....	169
Table 52: Experience in post-editing.....	170
Table 53: Estimated post-editing work in the last three years.....	171
Table 54: Different tasks at work	171
Table 55: Estimated daily throughput when translating from scratch.....	172
Table 56: Estimated typing speed.....	172
Table 57: Processing speed vs. Fuzzy match	179
Table 58: Processing speed vs. MT match	180
Table 59: Processing speed vs. No match	182
Table 60: Clusters and Speed groups	183
Table 61: Estimated mean per Cluster.....	184
Table 62: Estimated mean according to Match and Cluster.....	185
Table 63: Total errors for Fuzzy match in clusters.....	188
Table 64: Total errors for MT match in clusters	190
Table 65: Total errors for No match in clusters.....	192
Table 66: Estimated mean of errors per Match categories in clusters.....	193
Table 67: Estimated mean of errors in cluster.....	194
Table 68: Estimated mean of errors per match and cluster	194
Table 69: Question 1: Revision procedures.....	199
Table 70: Question 2: Post-editing learning curve	200
Table 71: Question 3: Post-editing proficiency.....	200
Table 72: Question 4: Post-editing effort	201
Table 73: Question 5: Price satisfaction	201
Table 74: Question 6: Job satisfaction.....	202
Table 75: Question 7: TM pricing	203
Table 76: Question 8: Fuzzy match revision	204
Table 77: Question 9: MT pricing	204
Table 78: Question 10: Ideal payment method for MT output.....	205
Table 79: Question 11: Predisposition to MT	206

Table 80: Question 1: Opinions of the on-line post-editing tool.....	209
Table 81: Question 2: Similarity with other tools	209
Table 82: Question 3: Usefulness of MT matches	210
Table 83: Question 4: Review method	210
Table 84: Question 5: Perceived productivity	211
Table 85: Question 6: Terminology	211
Table 86: Pricing for fuzzy matches a percentage of full word rate.....	268

List of figures

Figure 1: Screen-shot of the MT evaluation tool used	53
Figure 2: Human evaluation of MT output.....	54
Figure 3: Segments exported in Slicer.....	56
Figure 4: No match histogram	57
Figure 5: Dataset creation.....	58
Figure 6: Screen-shot of post-editing user interface used	59
Figure 7: Global processing speed according to Match category.....	88
Figure 8 Estimated mean of speed per translator	96
Figure 9: Sample for TR3 WPM	101
Figure 10: Sample for TR3 Standardized WPM	101
Figure 11: Sample for TR13 WPM	102
Figure 12: Sample for TR13 Standardized WPM	102
Figure 13: Correlation between TER, and Olivier and Hand	107
Figure 14: Correlation between TER and Levenshtein	107
Figure 15: TER indicator for all Fuzzy and MT matches.....	108
Figure 16: TER indicator per translator.....	108
Figure 17: TER indicator for Fuzzy and MT matches.....	109
Figure 18: TER score for Fuzzy and MT matches	110
Figure 19: Estimated TER per Translator.....	112
Figure 20: TER and Words per minute	118
Figure 21: TER vs. Categorized Words per minute	119
Figure 22: Estimated Mean of TER and Group Velocity	120

Figure 23: Overall TER values per translator.....	122
Figure 24: Total reviewing time	129
Figure 25: Total reviewing time in words per minute	131
Figure 26: Total time, in words per minute, taken by Reviewers	135
Figure 27: Total Errors Reviewers	135
Figure 28: Sample 1 correction Reviewer 1	138
Figure 29: Sample 1 correction Reviewer 2	138
Figure 30: Sample 1 correction Reviewer 3	139
Figure 31: Sample 2 correction Reviewer 1	140
Figure 32: Sample 2 correction Reviewer 2	140
Figure 33: Sample 2 correction Reviewer 3	141
Figure 34: Global error database vs. Segment error database	143
Figure 35: Error indicator per category	145
Figure 36: Total errors according to Match category	149
Figure 37: Classification of errors	153
Figure 38: Classification of errors II	153
Figure 39: Speed group and number of errors	158
Figure 40: Sample of biplot.....	174
Figure 41: Dendogram of clusters	174
Figure 42: Processing speed vs. Fuzzy match	178
Figure 43: Processing speed vs. MT match.....	180
Figure 44: Processing speed vs. No match	182
Figure 45: Estimated mean of speed per Cluster and Match.....	185
Figure 46: Total errors for Fuzzy match in clusters	187
Figure 47: Total errors for MT match in clusters	189
Figure 48: Total errors for No match in clusters	191
Figure 49: Estimated mean of errors and clusters	195

List of acronyms

BLEU	BiLingual Evaluation Understudy
CAT	Computer Assisted Tool
CI	Confidence Interval
DB	Database
GTM	General Text Matcher
LISA	Localization Industry Standards Association
LSP	Language Service Provider
Max	Maximum value
Min	Minimum value
MT	Machine Translation
NIST	National Institute of Standards and Technology
PE	Post-editing
Q	Quartile
Rev	Reviewer
SD	Standard Deviation
SWPM	Standardized Words per Minute
TAUS	Translation Automaton User Society
TER	Translation Edit Rate
TM	Translation Memory
TR	Translator
VM	Vendor Management
WPM	Words per Minute

PART I: Introduction

The world changes constantly and human beings need to adapt to this ever changing reality. Often this constant adaptation is challenging. There might be a temptation to stick one's head into the sand. However, change will still happen, with or without our participation and subsequent contribution of the reality ahead. In translation the same thing occurs: translation changes together with the evolution of language and technology. Localization, the area of translation that deals primarily with software and technical texts, has changed substantially in the last few years, as have other aspects of modern life. Often, economics sets the pace for these changes in localization processes. The new global reality means more volumes to publish, more languages to publish from, but – ultimately – the driving force is the desire to reach new markets and sell more products to increase revenues for shareholders. This is not a thesis about human economics or politics, but it is important to highlight that the interest in studying machine translation and post-editing comes from the realization of a fundamental change in the localization industry, which is in turn a consequence of economic changes produced in the world. Everything is, after all, interconnected.

In this context, machine translation is brought into the localization industry to minimize costs through the automation of the translation processes. This is done at this time because there is technology that allows it. If instead of machine translation, the localization industry could avail itself of another medium to lower costs, it would probably use it. Therefore, it is also crucial to understand how this affects the work of a translator, not only in terms of translation as a process but also its financial impact. After all, localization pricing has remained stagnant or has decreased over the last ten years, at least. If translators are going to be paid for machine translation post-editing, it is important that figures and calculations reflect the reality of the work, and not only be dictated by the need to lower costs by those that control the market.

New questions have emerged as machine translation (MT) technology has been implemented: How should MT segments be charged and paid? How much time would a translator take to complete the task of post-editing? How should this task be scheduled? What is the corresponding fuzzy-match value for MT segments? Should the same localizers be used or is there a new profile needed? Should these segments be reviewed

afterwards or is there no need to review them at all? Would quality be worse if translators used machine translated texts? Although there is more empirical data related to projects carried out with MT technology, there is no agreement on how it impacts on productivity, quality or the exact experience required to post-edit in a commercial context. There is also little empirical research on how translators behave or what they think about this technology. There is a lot to learn and a lot to be shared among the translation community: computational linguists, translators, agencies, buyers and users in general. With this view in mind, we decided to start this project: to learn more about the impact of machine translation in the workflow, primarily from and for the translators. This is the group of people that need to be involved and also be heard in the decisions that affect the present and future of their livelihood. If the world is changing, then everyone should contribute and shape it in a way that will allow everyone to thrive.

Research questions

The introduction poses several questions about the implementation of machine translation in the localization workflow. Since one single research project cannot aim to clarify them all, this research will be focused on the following questions:

RQ1: What is the corresponding TM fuzzy match price value for MT match segments according to the productivity obtained by translators?

By establishing a correlation between TM fuzzy matches and MT segments, more information about pricing can be gathered.

RQ2: Will using MT output have an impact on the final quality of the target texts?

By looking at the final quality of the target texts processed with and without machine translation, it can be established if using MT as part of the localization workflow impacts on quality.

RQ3: Is the translator's experience influential in the post-editing of MT output if speed and errors are considered?

By defining the translator's experience in relation to post-editing, we can explore whether experience has a role in the productivity and quality of the final texts, thus defining more precisely a profile for post-editors.

Thesis structure

This thesis is structured in four parts. "Part I: Introduction" includes Chapter 1, 2 and 3. Chapter 1 contains the research questions and an overview of Machine Translation and Translation Memory technologies. Chapter 2 contains a literature review of post-editing and translation memories. Chapter 3 presents the defined hypotheses, variables and operationalization, the methodology applied, and an overview of the project development.

"Part II: Quantitative results" includes Chapters 4, 5 and 6. Chapter 4 offers the quantitative results obtained for productivity, Chapter 5 presents the results obtained for quality, and Chapter 6 shows the quantitative results obtained for experience. Each chapter offers a brief conclusion of the findings in that particular area.

“Part III: Qualitative results” includes Chapters 7 and 8. Chapter 7 offers the qualitative results obtained from the questionnaires; Chapter 8 offers the qualitative results obtained from the debriefings with translators and a brief summary of the findings from both activities.

“Part IV: Conclusions” includes Chapter 9. This chapter offers the final conclusions, integrating both quantitative and qualitative results as well as lines for further research, and the reference.

Chapter 1: Overview of TM and MT technologies

This chapter gives a brief description of translation-memory and machine-translation technologies.

1.1. Machine translation and translation memory systems

In this section we will briefly describe machine translation (MT) and translation memory (TM) systems, what they are and how they work. It is not our intention to go through a detailed history or overview of these two translation aids, as we think they have been sufficiently explained and described in previous studies (Arnold 1994, Hutchins 1995, 2001, Austermühl 2001, Somers 2003, Bowker 2005, Gow 2003, Dragsted 2004, Quah 2006, and Ribas 2007, just to mention but a few) and by the software developers themselves (Systran 2012, SDL 2012, Star Transit 2012, Atril 2012, among others). We will give a brief definition of the tools and important aspects to consider for this particular study.

1.1.1. Machine translation

The definition of machine translation on the homepage of the European Association of Machine Translation (EAMT) reads:

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available which produce output which, if not perfect, is of sufficient quality to be useful in a number of specific domains. (EAMT 2012)

Although the definition is broad, since computers are used to translate texts in other forms that are not called machine translation, such as translation memories, it reflects the use of MT today. MT should be “useful in a number of specific domains” but not necessarily a replacement for human translation. The idea of a fully automatic high quality translation (FAHQT) has been replaced by a more practical use of human-aided machine translation (HAMT) within restricted environments.

As Bennett and Gerber (2003) explain, there are basically two types of MT engines. The first one, the rule-based system, maps semantic and syntactic structures from the source language to those of the target language. This is the engine traditionally used by large public organisms, such as the European Commission and the Pan American Health Organization, and also by pioneer companies in MT usage, such as SAP. It is still the preferred choice when available bilingual corpora are not extensive and for minority languages. This engine requires substantial initial work on defining rules, and additional time to maintain and update them. Examples of this type of engines are Systran, Lucy LT and Apertium. The second type of machine translation, the data-driven system, applies learning algorithms to large corpora of bilingual translations (once aligned, these are called “parallel corpora”), extracting translation parameters and models in order to find accurate translations for new material, basing the selection primarily on word frequency and word combination. A very large corpus is needed for the engine to be effective. This category includes statistical-based and example-based MT. Examples of this type of engines are Moses, SDL Language Weaver, Microsoft Translator and Google Translate.

Machine translation is used in different industries more or less successfully, especially in those that produce large-scale content of a highly repetitive nature (as is the content in the localization industry) that can be easily “processed” by an engine. MT is frequently associated with controlled language because if technical writers of source texts follow repetitive syntactic patterns, they will facilitate the implementation of MT solutions in a given company, thus increasing their translation capacity and saving costs. Even in this case, not everything is automatic in MT; there is a need for human interaction either before or after the machine has processed the data. The intervention before the machine processes the data is called “pre-editing”. It occurs at the source-language level to change language structures so that the machine translation engine is not confronted with ambiguous text. The intervention after the machine processes the data is called “post-editing” and it occurs at the target-language level to correct errors in the machine-translated output. To post-edit is defined as “to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)” (Allen 2003: 296). Post-editing is currently still essential to produce a high-quality product, meaning a product without frequent language errors as found in the machine-translated output.

1.1.2. Translation memory systems

Translation memory systems are widely used by agencies and freelance translators working in localization. Lagoudaki in a survey published in 2006 established that 82.5 percent of the people (interviewed) in this industry (translators, project managers, reviewers, terminologists and other translation professionals) use some form of TM system.

Translation memory systems are repositories or databases of human translated content. The systems store bilingual texts, also called bi-texts, in the form of sentences or paragraphs so they can be retrieved at a later stage and re-used if the new source text is identical (full match) or very similar (fuzzy match) to the previous source. When a new translation or an update is created, the system compares the new and old source, and finds an exact or similar correspondence. The proposed text is identified by a match or concordance level. For example, if a new sentence is said to be an “80 percent match” of an existing sentence, this means that the resemblance is high and only a few corrections to the target text are required by the translator. If the new sentence is said to be “100 percent match”, this means that there is a high probability of no change at all in the target text. The matching algorithms in TM systems are different for different products. The algorithm works on the syntactic structure of the source text and not on the actual meaning of the comparable sentences. As Somers explains, the algorithm works on “character string similarity that uses the well-established concept of ‘sequence comparison’, also known as the ‘string edit distance’ because of its used in spell-checkers, or more formally the ‘Levenshtein distance’” (Somers 2003: 38). Basically, this algorithm calculates the minimum number of changes (insertion, deletion, and substitution) that need to be made in one sentence (old source) so it is identical to the new sentence (new source), retrieving the closest correspondence in the target language according to this similarity.

Moreover, translation memory systems offer the translator concordance functions and terminology look-up features. These functions allow the translator to look for the translation of a single or multi-word term or phrase in the whole translation memory, and to paste the actual translation into the given segment.

All of these features contribute to productivity gains in the actual translation process, especially if the content is very similar to existing content or is highly repetitive within itself (as in the case of training materials produced in Word format and

almost identical content in PowerPoint format, for example). This productivity gain is taken into account when devising the schedule in a translation project or when deciding on the pricing structure (see Table 86 in Appendix A). Still, productivity is often measured in absolute terms, without considering the amount of work necessary to fix the errors and mistakes that might be a result of using the translation aid. Quality is expected to be the same as that produced by human translators.

According to the survey carried out by Lagoudaki (2006), the most widely used translation memory systems are (in this order) Trados, Wordfast, Déjà Vu, SDLX, Star Transit, Alchemy Catalyst, Omega-T, Logoport and Passolo (now acquired by SDL). This order may have changed in recent years.

Chapter 2: Literature review

In this chapter, we will present a summarized account of early experiments in post-editing and we will examine in more detail recent studies that are significant for this particular project. We will also give an account of the most recent empirical research on translation memories. There are other studies in the field of computational linguistics that are of interest for the development of machine translation and that have an impact on the future of post-editing, but they might not deal with or draw conclusions from the translators' perspective, which is the area of interest of this present project. We have included here only those studies that are closely related to post-editing. There are also other studies on translation processes, general reflections on the impact of technologies or summaries of recent changes in the translation field or training that, although of utmost interest, are outside the scope of the present study, and therefore are not included in this review.

2.1. Post-editing

2.1.1. *Early studies in post-editing*

Until very recently, little empirical research had been published in the specific area of productivity and final quality in post-editing. As Allen (2003) pointed out much information about post-editing is company specific, since companies carry out their own internal research according to their own preferred engine, processes, post-editing guidelines, desired final target text quality and other variables. In our own experience, the studies carried out within companies do not always have an appropriate scientific approach. Metrics are taken from real-life projects with different texts for different post-editors that contain MT, TM and new text. There is no tool to measure time, which means that the company has to rely on time measurements provided by each individual translator, many of whom are working from home.

Nonetheless, there were a number of pioneer articles in the 1980s and 1990s with information on MT implementation in different organizations, such as the European Commission and the Pan American Health Organization, describing the different tasks, processes and profiles in post-editing (Vasconcellos and León 1985, Wagner 1985, 1987, Vasconcellos 1986, 1989, 1992, 1993, Senez 1998) and specifying the different

levels of post-editing: rapid and conventional (Loffler-Laurain 1983, 1986). These articles touch on many fundamental issues related to the implementation of MT in a conventional translation process. Further, they give some indication of the productivity gains under certain restricted workflows. As far as we can see, on many occasions productivity guidelines are constructed based on average translators' metrics within the organizations, that is, the amount of words a translator might do in one day's work, but are not based on real and specific measurements in relation to human translation.

In the domain of machine translation used together with translation memories, there is an interesting case study exploring the translation of a help system from English into German at Baan (Andrés-Lange and Bennett 2000). Andrés-Lange and Bennett use Logos as the translation engine and Star Transit as the machine translation system in an innovative way: they integrate MT output into the Star Transit workflow process, treating MT segments as if they were fuzzy matches. They discover that they can reduce throughput times in this language combination by 50 percent, if the conditions are adequate, and they observe that MT productivity is in some cases lower than that of human translation. They highlight as well the importance of the human factor in the integration of TM and MT placing key importance on translator training.

Bruckner and Plitt (2001) present a methodology to use for the integration of machine translation output as translation memory input. The researchers define an evaluation procedure based on the ISLE taxonomy (created to evaluate machine translation quality). They propose that an evaluation scenario should consider speed, quality and user acceptance (usually translators). They also try to establish by means of an experiment a method for finding suitable TM fuzzy match values for MT segments. For their experiment they used software documentation in English translated into German by two teams of professional translators. One group translated using only TM while the other used a TM with MT segments inserted. The systems chosen were SDL Trados and a customized Systran system. The researchers conclude that in the case of a 76-percent fuzzy match the effort required, measured according to the edits made in both types of texts, for post-editing TM is less than the corresponding MT matches. Although this study uses very few sentences, the approach and the methodology could be adequate when trying to establish correlations between MT and TM segments. It further highlights the importance of determining the value for MT segments "in order to rank them adequately against fuzzy matches and provide the user with the translation candidate that requires least post-editing efforts" (2001: 5).

Possibly the most extensive research published on post-editing to date is Krings (2001) as part of his 1994 postdoctoral thesis. This is an extensive study employing Think Aloud Protocols (TAPs), mainly focusing on the mental processes used in machine translation post-editing. Krings looks at students translating instructional texts from English into German and French. He uses two rule-based MT engines (Systran and Metal) in what he calls a “black-box” model, that is, his aim is not to study the way the engines work, but to analyze the output as it is produced. He similarly tests post-editing with and without reference to source text, on-screen and on-paper. He measures temporal, cognitive and technical post-editing efforts. From his point of view, the decisive underlying factor determining temporal post-editing effort is the cognitive effort of post-editing. In his study Krings produces data on the processing time in seconds by task and processing speed calculated in words per minute, and he establishes definitions for absolute post-editing, relative post-editing and translation time. Relative post-editing is the relationship in terms of productivity between post-editing time and human translation. He concludes that post-editing machine translation generates very little time saving with regard to human translation. He estimates that the processing speed for post-editing on paper saves around 7 percent if compared to human translation, and post-editing on screen increases processing speed sometimes by 20 percent. This striking difference between post-editing on paper and on screen does not consider the time necessary to implement the changes that are produced on paper. Moreover, Krings states in his book that “Thinking Aloud as a data collection procedure has a clear slow-down effect on performance of basic tasks” (2001: 532). Krings not only concentrate on post-editing efforts, he also covers quality of the machine translations, description of the post-editing processes, quality of the post-editing material as well as extensive data on Think Aloud Protocols. He does not provide detailed information on times per post-editor but rather on the cognitive processes through verbalization techniques. All the same, Krings is a source of valuable information about post-editing and translation processes in general, although some of his findings on post-editing in particular are becoming outdated.

Jeff Allen has written several articles on machine translation and post-editing (2001, 2003, 2004a, 2004b, 2005a and 2005b) to spread information about this task and the use of machine translation within the industry and the translation processes. He has offered sound advice on the conditions required to acquire machine translation software and how to go about implementing it in a company. He has defined and unified different

concepts related to post-editing and its different levels (rapid, minimal and full post-editing) and established guidelines and criteria for this task. In the book edited by Somers (2003), Allen dedicates one entire chapter to describe the then largely unknown task of post-editing, profile of post-editors, and the usefulness of machine translation, and offers case studies from different organizations that use and create guidelines for post-editing. Further, Allen has also been involved in the development of a prototype automated post-editing (APE) module (Hogan and Allen 2000) in order to reduce the number of post-editing fixes that post-editors have to carry out on the actual output. Allen (2001, 2004a and 2005b) has conducted several tests on productivity, and offers some statistics on post-editing that show a pronounced increase in the productivity of translators when post-editing is compared with human translation. The tests calculate the total time used to post-edit and review, including the time used to create the dictionary in a rule-based MT system and the output from machine translation in comparison to a fixed number of words per day from a human translator (2,400 words per day).

Following in Allen's footsteps, Lorena Guerra wrote a dissertation on the full post-editing of raw machine translation output of marketing texts (Guerra 2003). She uses the PROMPT engine (a rule-based machine translation engine) to translate three marketing brochures from English into Spanish. The project is performed by two participants: one carries out pure human translation and the other, Guerra herself, uses machine translation and then post-edits the texts. She provides the times taken to read, research terminology, translate and review by the human translator, and to read, automatically translate without dictionary, identify unknown and mistranslated terms, code entries in a user dictionary, reprocess with automatic translation, post-edit and review the same text by the post-editor. She observes an overall increase in productivity of the translation cycle when using machine translation. According to Guerra, companies can attain almost triple the savings if machine translation is used. Like Allen, she uses an average translation rate of 2,400 words per day as a guide for human translation. Guerra presupposes that the final quality of the samples is acceptable because all samples were translated, either automatically or humanly, and reviewed, but she provides no check on the final comparative qualities of the outputs.

The ERP software developer SAP has been a pioneer in the use of machine translation and post-editing raw output within the localization industry. Schäffer (2003) describes a project carried out at SAP in conjunction with their language-service

providers to put together common practices and guidelines for post-editing, thus making the task more accessible to translators, as well as creating a general typology of MT errors. This error classification would make the post-editors aware of the most common corrections. Several translation engines were used (Logos, PROMPT, Metal and Logovista) all of which are rule-based machine translation engines, and several language combinations. Schäffer concludes by highlighting the importance of controlled language input in order to maximize the productivity of MT output. Nevertheless, we find no data on productivity measurements or quality of the final target texts.

2.1.2. Post-editing and cognitive effort

Sharon O'Brien has studied various aspects of post-editing: post-editors' profiles and course contents for post-editing (2002), and the correlation between post-editing effort and machine translatability, suggesting a new methodology for measuring source-text difficulty and cognitive effort (2006a). Very relevant to this study, though, is her research on eye-tracking and translation memory matches (2006b). In this paper, O'Brien analyzes eye-tracking as a methodology for recording a translator's interaction with translation technology and explains differences in cognitive effort with different translation memory match types. She is interested in the "cognitive effort required from translators for different match types in a TM system, with a particular emphasis on comparing matches generated by MT systems with other match types" (2006b: 187). This is quite relevant because she is not focusing on comparing actual human translation with MT but she is establishing a relation between MT and fuzzy matches in TMs. In fact, most uses of MT are similar to the uses of TMs: both "techniques" are regarded as aids to the translator. As O'Brien mentions, "we are interested in this question because, in industrial settings, assumptions are made about the effort required to process MT matches (and the amount a translator should consequently be paid), without empirical investigation" (2006b: 187). In her study, O'Brien has four participants familiar with SDL Trados translate a text from English into German and French using different match types, fuzzy matches, and introducing some matches from MT using Systran as the engine. The result is that exact matches (100 percent matches) present the least cognitive effort and no matches demand the greatest load. Furthermore, O'Brien shows that as the Fuzzy match value decreases, the cognitive load increases and that MT matches appear to be equivalent to 80-90 percent Fuzzy matches in TMs in terms of

cognitive load. She points out that these findings would need to be validated with a larger number of participants and segments.

2.1.3. Post-editing and quality

Bowker and Ehgoetz (2007) explore user acceptance of machine translation output. They carry out an experiment where they present three different target texts for the same source text (French to English): one human translation, one raw MT output and one post-edited MT output to 121 professors at the Arts Faculty in University of Ottawa. They judge these documents according to speed, quality and cost. The researchers are trying to determine if these users would accept lower quality in exchange for lower cost and faster turn-around times, or whether they would prefer higher quality at a higher cost and slower turn-around time. The results show that two thirds of the participants (21) choose the post-edited option and one third (10) the human translation. However, the study takes a standard measurements for human translation (2,000 words) in contrast to real post-editing times, it considers only the time for processing the output in a rule based machine translation system (Systrans) and not the time needed to set this system up or update it, and the translations costs are calculated making certain assumptions about percentages (which might not correspond to the localization industry). The study is nevertheless innovative because it involves the recipients of the translated documents. In other words, it involves “the user”, and points to the idea that language professional participants are more linguistically sensitive to language quality than those that are not language professionals.

Fiederer and O’Brien (2009) examine the question of quality in machine-translated texts. Their premise is that professionals would tend to think that the raw output post-edited by translators would show a lower quality than the texts translated by human translations. In order to test this, they setup an experiment where eleven raters evaluated 30 source sentences, three translated versions and three post-edited versions (180 sentences in total), according to the clarity, accuracy and style parameters. The raters were asked then to apply a four-point scale to each sentence going from 1 to 4 (although different for each category, 1 represented a low mark while 4 represented the highest). They were also asked to indicate their favorite translated option out of the six proposals for that given source (they were not aware that they had to rate post-edited text). The raters gave equal scores for translated and post-edited sentences with regards to Clarity, higher scores for post-edited sentences with regards to Accuracy, and finally

they gave a higher scores to translated sentences in terms of Style (in this category the difference was greater between translators and post-editors, but still both were closer to the 3 point scale, and post-editors had been instructed not to change the text just for the sake of improving style). Further, raters chose primarily the translated sentences as their favorite sentences (63 percent of the sentences as opposed to 37 percent for post-edited sentences). The study also finds possible correlations between the use of controlled language rules and the quality of the post-edited product. This study is interesting because it is not frequent that quality is analyzed in post-editing research or simply commercial experiments. More focus is placed on the productivity of the translators or post-editors than on the final product. It is important to note, however, that in this instance translated sentences were not reviewed, so, the final quality produced might not necessarily be that of a “final translated product”, while post-editors had “the advantage” of starting with a previous translation of some sort (although it was MT output, it had been generated by human-translated bilingual corpora). Post-editors also had the disadvantages that this type of translation might represent in terms of time and style, perhaps if different instructions were given to post-editors, they would have modified the style and therefore achieve a better final result in the style category from the raters. Having said this, we understand that this study intended to explore the idea that translators “on their own” produce better translations than those combining machine translation and post-editing.

Carl et al. (2011) compare the post-editing experience in a group of translation students and professionals, using Translog to measure time, keystroke and gaze data. The researchers use three texts from English to Danish with an average of 850 characters each to translate manually, at the same time the same texts were machine translated using Google Translate. The quality of the translations was evaluated by seven native Danish speakers (four were professional translators). Each evaluator was presented with four translations, two manual translations and two post-edited texts and they were asked to rank the translations. The results show that the post-edited texts are judged to be better than the equivalent manual translations, although this difference is not statistically significant. The edit distance is calculated for the post-edited segments and compared to their quality ranking, and there is no correlation between the number of edits and the score, indicating that more post-editing does not necessarily lead to better translations. As to time, post-editing is done faster than manual translation but the difference again is not statistically significant possibly because the volume of work

process is low (very few sentences with low word-counts processed in over seven minutes). When gaze data is analyzed, the results show similar times in both activities. However, in manual translation more effort is placed on the source text, while in the post-edited text, the target text requires more consultation. The researchers suggest that the post-editor would consult the proposed text first and then look at the source, while in manual translations it is obligatory to read the source text in the first place. Although, we believe the methodology for this experiment attends to several aspects of post-editing that need addressing (speed, quality of post-edited text, cognitive effort), the actual experiment presents several problematic areas: the volume is low, the profile of the participants (both reviewers and translators) is rather mixed (some of them participated in both the manual and the post-edited version with a time gap), the nature of the text is general (this could be beneficial for an engine such as Google Translate) and the quality of the MT output is unknown.

García has written several articles in relation to translation memory technology and research (2005, 2006a, 2006b, 2007, 2008, 2009). Most recently, however, García (2010) has explored the use of machine translation and post-editing in a non-professional context. A test with fourteen educated bilinguals with an interest in translation is setup from English into Chinese using Google Translator Toolkit (GTT) in order to assess if the quality and speed of texts translated with the automatic machine translation option in GTT is higher or lower than without the machine translation option selected. The participants translate four 250 word extracts (that is 1000 words in total) from general texts dealing with legal and medical topics. Two texts are translated with MT and the other two from the source text directly. Two markers assess the quality using the guidelines from the Australian National Accreditation Authority for Translators and Interpreters (NAATI) and without being aware of the actual origin of the text before translation or post-editing. The results show that translating with MT is faster in 15 cases out of 28 cases. There are, however, no statistically significant differences between the two groups. Regarding quality, the MT solution is preferred in 59 percent of the cases. This difference is statistically significant for the two groups for Evaluator A, but not so for Evaluator B. Further, the researcher does not find data to support the hypothesis that “poor” performers did better using MT than just working from the source text. Although this experiment is not performed with professional translators, the engine is not trained specifically with domain data, and time measurement is not explained, it is interesting to see that independent evaluators rated

the post-edited segments higher than those translated without MT. It is important to point out, however, that the translators were not allowed any reference material (in order to avoid the impact of other variables such as search strategies) and this might have played a fundamental role in the final quality, thus favoring the MT post-edited segments.

García (2011) setup a second phase of the previous study, with 14 students from English to Chinese, and a further third phase with 21 students from Chinese into English (they were to process 500 words in approximately two hours). The participants are divided into two groups. One control group translates using GTT without any suggestions and another experimental group translates using the MT baseline. Translators are requested to translate to “full quality” (ibid: 220) and this is measured using the NAATI framework, as in the previous experiment. The results for the second phase show, as they did above, that post-editing does increase speed but this increase is not statistically significant in relation to translation without MT. García also identifies passages in terms of difficulty for translators. For the easiest passages, post-editing is always faster. For the most difficult passages, translation is faster than post-editing. As for quality evaluators, they seem to show a “big disparity” (ibid: 224) but still post-editing scores higher than translation (an average of 36 while translation scores an average of 34). The results of the third phase show that there is an increase in speed when post-editing and, in this case, the increase is statistically significant. Furthermore, if the easiest and the most difficult passages are considered, post-editing is always faster than translation. Again, quality (judged by only one marker) was higher in post-edited passages. García also examines best and worst performers, according to the marks received, and the results show that both groups perform as well when post-editing as when translating if speed is considered, and both perform better when post-editing if quality is considered, “although we could infer from the figures that weak performers benefit more from post-editing than the stronger ones when translating into their mother tongue” (ibid: 227). We have not seen any prior test dealing with post-editing into a second language and the results here are very positive. It seems as if MT can help at least students or even “amateurs” to produce better texts in a second language. It might also be surprising for some that post-editing scores better in terms of quality than does “normal” translation. As García points out:

In contrast, our results seem to fit with other recent research which also supports the notion that post-editing can be an appropriate alternative to conventional translation rather than a second-best solution, which while faster is of lower quality. (ibid: 228)

De Sutter and Depraetere (2012) present a study with 15 translation trainees, two of which are non-native speakers, who post-edit and translate 3045 words from English into French (half of the text is post-edited and the other half is translated from scratch). The MT output used was from a customized RBMT system (dictionary coding of 22 words). The post-edited text is evaluated by a professional translator using a five point scale. The post-editors used the company CrossLang's on-line post-editing tool to process the segments. The results show that the average productivity increase for translators when post-editing is over 22 percent and over 30 percent if the two non-native speakers are included. The standard deviation, however, is high among the participants, the productivity increase ranges from 1 to 91 percent. When looking at the quality, the manual translation score is slightly higher than the post-edited version. However, the scores are high for both modalities, and the edit distances between the two translations are similar. Despite having translation trainees instead of professional translators, this study presents interesting results and a methodology that is in line with the methodology used in our research project.

2.1.4. Post-editing and experience

De Almeida and O'Brien (2010) explore the possible correlation between post-editing performance and years of translation experience. This pilot experiment is carried out with a group of six professional translators (three French and three Spanish) in a live localization project using Idiom Workbench as the translation tool and Language Weaver as the MT engine. The file to translate is from the IT domain and it contains 350 words. The Language Weaver engine had been previously trained for both languages with approximately 3 million words. Since the intention was to measure the post-editing performance in relation to experience, four translators had experience in post-editing while two others did not. To analyze this performance a LISA QA Model is used in combination with the GALE post-editing guidelines as the former is not deemed suitable on its own to measure the post-editing activities since it was created to evaluate human translation. The researchers divide the correction of errors into essential changes (errors in MT output that really ought to be fixed according to the guidelines given) and

preferential changes (edits that are deemed to be unnecessary according to the post-editing guidelines). The results show that the translators with the most experience are the fastest post-editors and they make the higher number of essential changes. However, the results also show that the translators with more experience make more preferential changes. The methodology for this study is interesting for us, especially the edit analysis divided into different categories (preferential and essential) and in relation to experience. However, the sample here is too small to see trends, we understand this is part of a larger project, but still with six translators it is difficult to know if these observations are particular to this group of translators only. Also, we feel that when analyzing error typology during post-editing (for example, if more Language errors are found in this particular case) it is important to look at the errors when these same post-editors are translating on their own as the reason for having more Language errors, for example, might be that the translators make certain mistakes regardless of whether they are translating or post-editing. Finally, it is unclear in this article who is classifying the errors, and the reviewer might also be influential in the results.

Depraetere (2010) analyzes text post-edited by ten translation trainees (Master students) in order to establish post-editing guidelines for translators' post-editing training. The students post-edit 2,230 words of support options from a Sun operating system. The translation is divided into two sets of 55 segments. The first half was pre-translated using a rule-based MT system and the second half with a statistical system, both customized. Each segment was translated by the ten students using a CrossLang web-based post-editing tool. The students received minimal instructions because the researcher was interested in knowing what the students would correct intuitively: to make sure that the source and the target texts have the same information and that the target text is grammatically correct, with a few examples of necessary and unnecessary post-editing changes. The analysis shows that students follow the instructions given and they do not rephrase the text if the meaning is clear, the students "did not feel the urge to rewrite it" (ibid: 4), they are not, however, sufficiently critical of the content thus leaving errors that should be corrected according to the instructions. Depraetere points out that this indicates a "striking difference in the mindset between translation trainees and professionals" (ibid: 6). She also highlights the need to give very clear and detailed instructions on what exactly requires to be changed when using machine translation. Despite the fact that this study is focused on students, we find that it might be applicable to junior translators who have been exposed to machine translation either during their

training or from the beginning of their professional experience as opposed to more senior translators that might have experienced MT at a later stage in their professional life. There appears to be a new generation of translators that would need a set of instructions different from those that are currently given to professional translators.

2.1.5. Post-editing and automatic metric scores

Tatsumi (2009) explores the correlation between automatic metric scores and post-editing speed, she examines other factors such as segment length and structure, and dependency error. The objective is to find a quick and easy method to analyze post-editing (PE) effort overall. For this experiment, three Japanese translators process 4,784 words through SDL Trados Translator's Workbench with a Systrans engine for the machine translated output. Time was recorded using SDL Trados and a specific macro devised to achieve more thorough measures. The metrics used are: GTM (General Text Matcher) (Turian et al. 2003, Melamed et al. 2003), TER (Translation Edit Rate) (Snover et al. 2006), BLEU (BiLingual Evaluation Understudy) (Papineni et al. 2002), and NIST (National Institute of Standards and Technology) (Doddington 2002). The results show that GTM has a stronger correlation with post-editing speed. However, it varies depending on the sentence structure being stronger for simple sentences (containing only one clause) and weaker for sentences that are complex (containing one or more subordinates) and incomplete (incomplete if not observed within a context) sentences. The segment length also has an effect in PE speed: very short and very long sentences are slower to edit, and dependency errors have a greater impact on incomplete sentences. Overall, the relationship between automatic scores and PE speed is not a linear one, since source text characteristics and MT errors might affect the speed. Tatsumi concludes that cognitive effort and therefore the time spent solving the different issues might play a fundamental role on this correlation.

As a follow up to this study, Tatsumi and Roturier (2010) explore the relationship between text characteristics (ambiguity, complexity and style compliance) and technical and temporal effort. The objective of this research is to find characteristics in the text that affect post-editing with a view to ultimately designing tools that could inform the post-editor about the effort require for this particular segment. They carried out an experiment with nine translators from English to Japanese, using a Systran engine and Symantec corpus (3,916 words). The Systran engine calculates the text ambiguity and complexity using scores that evaluate the text from 1 to 4 (from less to more complex).

Ambiguity refers to single words with multiple meanings and complexity to the structure of the sentence. The software “After the deadline” and “acrolinx IQ”¹ was used for scores on style and grammar. The information given by these tools was measured against an automatic metric score, GTM, and the time post-editors spent doing the task was measured using SDL Trados Translator’s Workbench with a customized macro. The results show that there is a strong correlation between complexity and ambiguity scores and technical PE effort and a moderate correlation with acrolinx IQ scores and temporal PE effort. The researchers remark that “both within and between post-editor variance is much higher in terms of PE speed compared to the amount of technical PE effort” (ibid: 50). They conclude that the post-editors perform more or less edits (technical effort) in a given sentence due to this sentence ambiguity or complexity, but the time spent (temporal effort) cannot be directly linked to these scores.

O’Brien (2011) investigates the correlation between GTM as well as TER and post-editing productivity measured according to speed and cognitive effort (with an eye-tracking device). An English-to-French corpus was used containing 55,000 sentence pairs, 10,000 of which were reserved as a test set, and 45,000 were used to train a data driven MT engine. From the test set, 995 sentences are chosen randomly according to distribution of their GTM scores and organized as Low, Medium and High score categories. TER, on the other hand, is used as a post-task measurement with the idea of establishing correlations with productivity and GTM scores. Seven translators post-edited 782 source words into French. The tools used were Alchemy Catalyst as the post-editing environment and Tobii Studio as the time-recording and eye tracking tool. The results show that the GTM categorization and TER scores correlated well with processing speeds for groups of segments but not necessarily with individual segments. When looking at fixation time, the Low GTM category requires the most fixations per word (most effort) and the High GTM category, the least. The Medium category, both in processing speed and fixation time is closer to the Low than to the High category. As O’Brien points out, the correlations with these automatic scores and post-editing effort are useful to assess the actual effort required to post-edit but they cannot help commercial environments to determine how useful the MT output is or how much effort is required to post-edit a given sentence before the actual post-editing task occurs. It

¹ <http://afterthedecline.com/>
http://www.acrolinx.com/acrolinx_iq_en.html

can, however, be used as a predictor. We believe that because a post-editing task is not a once-off endeavor and presumably one post-editing task, for a particular customer using a particular engine in a language combination, leads to future similar tasks, calculating these scores in a first project might give an accurate estimation of the actual quality that MT output for future projects, especially if final quality of the post-edited segments is also factored in.

De Sutter (2012) examines the correlation between edit distance and fluency scores. She setup an experiment using 2,300 words from an English leaflet. The output from two engines (one RBMT and one SMT) was selected to create the evaluation sets for the post-editors. Eleven students from a Masters in Translation Studies were asked to evaluate the two sets from the two engines in a slightly different order. Both outputs were evaluated by ten informants. They evaluated the fluency and the accuracy of the output using the DARPA five point scale system. The results show that the RBMT system generates the best results according to the human informants and also requires the least amount of time to post-edit. De Sutter also looks at the edit distance between the output and the edited text, using an algorithm in PHP (Olivier and Hand 1996). The closer the value is to 100 percent the closer the strings are. The results also show that the RBMT system scores the lowest edit distance when comparing the output to the post-edited version, and the scores also show a correlation with the human quality rating. From these results, De Sutter proposes an edit distance mapping with human evaluation. The score of 100 corresponds to an Excellent Fluency score (5), the 95-99 to Good (4), the 80-94 to Average (3), the 50-79 to Poor (2), the <50 to Very poor (1).

All these studies highlight the need to establish a correlation between automatic scores and the post-editing effort, something that the localization industry needs in order to set a pricing scheme for a given MT project, but also to plan and organize complex projects using MT output and post-editing. For this reason, they are innovative and important in post-editing research.

2.1.6. Post-editing and confidence scores

In the field of computational linguistics there are some interesting studies that although not exactly related to this particular project, explore the integration of MT and TM and in doing so involve post-editors not only in the evaluation of segments but also as the object of the study. The idea behind the studies is to obtain confidence scores in the MT output that will help post-editors to know in advance which segments require a higher

degree of editing. The idea of confidence scores is not a new one and we do not intend to review all literature on this particular area of computational linguistics, but we will review recent studies that might indicate the way in which MT might be used with TMs in order to optimize post-editors' effort and final quality of the target texts.

Specia (2009a, 2009b) investigates the problem of predicting MT output quality when there are no human references available by applying regression estimation models to obtain scores in various MT systems and language pairs at the sentence level. The objective of these investigations is to facilitate post-editors the evaluation of MT output so they do not have to spend time deciding if the segment is of sufficient quality to be edited. Although these are preliminary studies, the results show that with this method it is possible to control expected precision and recall, and therefore select a set of translations that are very likely to be of better quality for the post-editors.

Specia (2011) experiments with different estimation models on three annotations types (post-editing time, distance and effort scores) in order to improve how different MT segments are flagged, that is, to give a confidence estimation or confidence score. As Specia remarks translators often complain that the "post-editing of certain segments with low quality can be frustrating and can require more effort than translating those segments from scratch, without the aid of an MT system" (ibid: 73). This can sometimes cast a shadow over those other segments that are of higher quality, and, evidently slow down translators' work. Specia setup an experiment with post-editors so they could evaluate the quality of the machine-translated segments from 1 to 4: 1 being a complete retranslation and 4 a perfect proposal. At the same time they were asked to post-edit in order to calculate the edit distance of each segment through an on-line tool that also measures the time it takes them to complete the post-editing of each segment. At the end of the task, Specia analyzes the data by means of establishing the quality of the raw output in relation to translators' opinions on effort, the edit distance and the time spent. The results show that "CE [Confidence Estimation] models that learnt from objective annotations of translation quality produce rankings of translations that reliably reflect their post-editing effort" (ibid: 79). Although this is a highly technical paper from a computational linguistic domain that is beyond our specific knowledge, we find that it addresses the need to obtain scores per segment similar to the ones offered by translation memories in fuzzy matches, and it involves post-editors as human annotators. This is extremely interesting, because in a setting where there is a MT

engine in place, the work done by post-editors can in turn generate confidence scores for future segments and inform about the quality of the MT output.

He et al. (2010a) propose a framework that integrates MT with TM for translators. The framework recommends MT outputs to a TM user when it finds that the MT output is more suitable than the fuzzy matches suggested by the TM. To evaluate it they use an automatic MT metric (TER) and obtain a precision of 0.85 and a recall of 0.89. Therefore, post-editors can continue to work in a TM environment including MT with the same pricing system as with TMs. This is relevant to our study as it uses TM to benchmark results from the MT output: if the effort (measured in terms of edits) is equal in both TM and MT segments proposed, then the pricing can be equal. Post-editors are not involved in evaluating both type of segments but this is outside the scope of this first study. We wonder, however, if number of edits (through TER) is a valid measurement when used in isolation, that is, if temporal effort is not considered.

Following up on the study above, and more interesting for us, He et al. (2010b) explore the integration of SMT systems and TM, to explore how this integration can help translators to choose the most effective option during the translation process through a recommendation model. In this case, rather than having an automatic metric, they conduct a human evaluation on TM and MT integration with a team of five post-editors from English to French through a web application. Post-editors choose the segment best suited for post-editing from a set of three segments on screen, the original text, and two target texts (TM or MT segments randomly placed on the screen) without knowing the origin of the given text (the two target texts are labeled as Candidate 1 and 2). The time employed in deciding is measured and a post-assignment questionnaire is sent to the translators to know more about their experience in post-editing and their opinions of MT. The results show that all post-editors selected more MT outputs to post-edit than the TM options. This supports the results obtained in the previous study with automatic evaluation metrics. They also note consistency in the user behavior according to inter- and intra-annotator agreement. During a post-assignment questionnaire, one post-editor said that the quality of the TM proposals was more suitable for post-editing when in fact he had chosen more MT outputs in the experiment. We can see here that as the quality of SMT engines improves, it is harder for post-editors to distinguish between the quality produced from the engine and from TMs, and that if the high quality MT segments are proposed, post-editors might even prefer to use MT output. Evidently, in this experiment post-editors do not actually post-edit the

proposals in order to see if they were faster when post-editing these segments than when translating or editing TMs. We understand that this would require a completely different setup as the same post-editors cannot evaluate and edit the same source sentence since this would alter the final results.

2.1.7. Post-editing in a commercial setting

Flournoy and Duran (2009) investigate whether post-editing MT output is faster than translating from scratch within the context of integrating MT in the Adobe production workflow. They carry out two tests, a small pilot of 800-2,000 words and a second larger project of around 200,000 words using two MT engines: PROMT for Russian and Language Weaver for Spanish and French trained with Adobe data and lexicons. The results show that there is an increase in speed when post-editing in both cases. In the small pilot the results show that a translator's daily output can be done in less than two hours when post-editing. As the researchers point out, these figures do not take into account any overhead for the project and are used in comparison to a standard figure of 2,500 words per day, not the real productivity of these translators. Although exact data are not provided for the second test, they report increases of 40 to 45 percent in speed in comparison to a human translation. It is difficult to know how these figures are calculated in the context of a live project. Although in the second stage real figures are measured (post-editing versus human translation) the files are different and as a consequence it is difficult to know if the speed is related to MT or to the nature of those specific files. This is one of the issues when measuring in live contexts: there are too many variables. There are, however, interesting comments in this paper: MT quality and editing speed vary significantly between files; MT quality was related to the quality of the source text (if the two tests are compared); the integration of a Globalization Management System (GMS) and MT requires thoughtful consideration before implementing MT in the regular translation process (otherwise benefits from translation memories might be overlooked); feedback from translators interferes with their speed and it should, therefore, be compensated; and post-editing requires senior and skilled translators because novice translators trusted MT output more readily than senior ones, and this can lead, presumably, to a lower final quality.

Groves and Schmidtke (2009) analyze post-editing patterns in MT projects carried out in Microsoft. The engine use in this case is the Microsoft Treelet system (Quirk et al. 2005) customized with their own data. Microsoft relies on the time

measures provided by their language service providers from three “representative” translators (one of average productivity, one new to the project and one expert translator). Translators log in their own time. The results obtained are then averaged to know the productivity gain per translator. Further, Microsoft gathers unstructured feedback from their translators (through the LSPs). Microsoft reports improvements in the quality of the MT and related productivity increases from 5-10 percent to 10 to 20 percent for certain languages, although they signal variations in post-editing productivity for the same language depending on project, product, different file deliveries of the same project, and between different translators. Translators report on issues related to terminology, grammar, and incorrect handling of mark-up and formatting (tagging). To analyze the post-editing patterns, two data sets are used: English into German and into French. The source text, the machine raw output and the post-edited text are used for the analysis. The distribution of segment length patterns is similar for both languages with a majority of segments containing 20 words or less. Using their own edit distance techniques, they find that for French the edit distance is 5.60 whereas the German score 8.81, indicating a greater post-editing effort for German. The most common types of edits are deletion and insertion of function words (especially determiners). There are also edits in punctuation, especially actions related to inserting or deleting commas. They also give a detailed report on structure- based comparison for each language. We think that the time measurements used for this study, and therefore the productivity gains reported, contain several weaknesses, translators measure their own time from home while working on a “normal” project, so the time measured does not reflect human translation in comparison to MT post-editing, but it includes different levels of recycling material (through TMs) and different type of files across translators with the same language and also with different languages. It is difficult to know if the productivity variations across translators or languages are due to the way the data is gathered or to the impact of MT in the localization process. The analysis of post-editing patterns is an interesting area, although in this case the analysis is merely descriptive.

Plitt and Masselot from Autodesk (2010) present a very interesting report on a productivity test of statistical machine translation post-editing. It is significant for us for two reasons: it is in line with our line of research (Guerberof 2008) both in methodology and findings, but furthermore the researchers work in a commercial setting. This is important because it signals, in our view, a slight change in how post-editing and post-editing “experiments” are being viewed and carried out in the localization industry.

Since the origin is commercial, the study is more focused on the productivity of the post-editing task, and this is understandable, since an increase in productivity can mean a reduction in costs, and therefore, more profits for the company. However, the advantage of these types of studies is that they are carried out in an environment “similar” (with similar texts) to that of professional translators. In this particular case, Plitt and Masselot setup an experiment with twelve participants translating from English into French, German, Italian and Spanish. They use a Moses (Koehn et al. 2007) engine trained with Autodesk data up to 2008 and then a random subset of their own content from 2009 (a total of 144,648 words of source words processed). The translators worked in a post-editing environment specifically created for the purpose of the experiment, containing the source and target fields exclusively. The Autodesk QA team verified the final quality produced by the translators. This team was aware of the productivity test but not of the particular origin of each segment. The results show high variance across translators; MT allowed translators to work faster but the percentages varied from 20 to 131 percent. They also find that the benefits of MT are greater for slower translators than for faster ones. Interestingly, the edit distance (that is, the number of edits done in one segment when post-editing) is not lower for faster translators: the fastest translator has the highest number of changes. Therefore, no correlation is found between edit distance and throughput. When looking at segment length, they point out that the optimum length for MT appears to be between 20 and 25 words: for smaller or larger segments the productivity is not as pronounced. The Autodesk QA team reports, after reviewing all proposals, a higher number of errors for translation jobs than for post-edited jobs in all the languages tested. There are other findings related to keyboard activity and pausing where MT seems to save not only on typing but also on “thinking” time, and the fact that MT evens out the work for translators. All of these are very relevant findings that could be to some extent verified in the present study. It has to be said that Autodesk has a very large bilingual corpora in all these languages and an established QA process for these memories, as well as extensive terminological databases. This increases the quality of the MT output. Finally, Autodesk has the financial means to test machine translation productivity and quality in a number of languages, with a high number of translators and an array of content that sometimes is not available in smaller companies. This is, in our opinion, the most thorough research carried out by the private sector that we have found to date.

With subsets of the data gathered from this experiment, Beinborn (2010) performed a cross-linguistic analysis of the temporal, technical and cognitive effort in the post-edited text (German, French, Italian and Spanish) as part of her Master's thesis, with the view of understanding better the post-editing process. The segments that required more time in the previous experiment are identified and technical annotations are applied to verify the type of edits made in the MT output. Also, the translators' feedback is analyzed to better understand their experience. Pause indicators (measured by the absence of keystrokes) are analyzed to measure cognitive effort. From a temporal perspective, the results show that the source segments that cause high temporal effort differ from language to language, although some short segments (without context) seem to cause similar temporal effort from language to language. From a technical perspective, the results show that a higher number of edits (using her own annotation system) per segment lead to a higher processing time. Across the four languages, the distribution of edits is similar, and four general source properties cause an increase in the technical effort: long segments, tags and technical instructions, and complex descriptions. From a cognitive perspective, the results are similar to the results for temporal effort results, since segments with a long overall duration also have a long pause time. The pause time measurement, however, has certain limitations (it includes both non-typing time and other actions such as mouse movement) and for this reason Beinborn indicates that more data is necessary for a cognitive analysis. She also finds that the target language plays an important role with regard to cognitive effort. Finally, the segments that cause an increase in pause time correlate in translation and post-editing. We are unsure about certain decisions made, for example normalizing the source segment length with the length of the raw MT output because translators work with this text "and only use the source text as a reference to check the intended meaning" (ibid: 35). This is an assumption that we do not think has been sufficiently tested. Also, the complex annotation scheme applied to the post-edited text to know the nature of the edits could have been complemented by using automatic scores (as in Tatsumi 2009). Finally, the methodology used to measure cognitive effort has shortcomings, as indicated by the researcher. Still, the decision to study the post-editing and translation process using a novel system of annotation in the target languages is of high value, and it remains an interesting model to follow if post-editing is to be analyzed in a detail manner at a target sentence level.

Tatsumi (2010) studies as part of her doctoral thesis the speed of nine professional English to Japanese post-editors and analyzes the amount of editing made during the process, as well as the influence of the source text on speed and type of edits. A 5,000 word corpus was extracted from a user manual from Symantec Corporation. The post-editors used SDL Trados Translator's Workbench with a specific macro designed to help to measure time. A small sample (906 words) of TM segments was included (from the 75 to 99 percent fuzzy match range). The MT engine used is Systran version 6 and 3 with pre-processing and post-processing scripts. The participants were asked to fill in a post-assignment questionnaire in order to understand the post-editors' differences that could influence the process and also to gather their opinions both about post-editing and machine translation in general, and about that particular project. The results show that the amount of editing moderately correlates with post-editing speed. This correlation is stronger for simple and complex sentences and weaker for incomplete sentences. The variance in speed within and between post-editors is higher when speed is examined than when the edit distance is measured using GTM scores. This means that post-editors with different speeds might make the same number of edits. The results also show that simple sentences are the fastest to post-edit, followed by complex ones and finally by incomplete sentences; procedural texts are faster to translate than other types of texts, and the more complex the sentence, the slower the post-editing speed. The PE operations that have an impact on speed are: supplementation, rewrite, and punctuation edits. The editing speed of Fuzzy matches (75-99 percent matches) is not significantly different from MT matches in terms of speed. However, Fuzzy matches require more lexical changes while MT requires more grammatical changes. Interestingly, when looking closely at the number of revisits in the same text, it is found that some of these revisits are not made to make a change but to review a previous decision or change. However, the post-editors that have a higher number of revisits are not necessarily slower. The opinions from the participants highlight a need to standardize terminology in the MT output, and they indicate a positive attitude to the quality and potential of MT output. We feel that the sample of participants is small if statistical conclusions in terms of speed are to be drawn. However, this is a very thorough and solid study that addresses important questions in post-editing through the behavior of post-editors in a commercial environment, and it sheds some light on the nature of changes and their correlation to speed.

Following up on the experiment carried out by Plitt and Masselot (2010), Autodesk (2011) has published on their website the results of a two-day translation and post-editing test carried out with 37 participants from English into Chinese, Japanese, Polish, Portuguese, German, Italian, Korean, Spanish and French. They used a Moses engine trained with Autodesk data, and a specific translator interface was used to record the time spent in each segment. The corpus is composed of user interface as well as documentation segments. The results show that for all languages and participants, post-editing productivity is higher than translation productivity. The productivity increase is different for different languages, for example Chinese is the lowest (with 42 percent) and French is the highest (with 131 percent). Spanish has a 117-percent increase. Experience, especially in post-editing, is considered to be the most important factor in productivity. Translators' perception of speed does not correspond to the actual speed they experienced during the test, especially those that believe they are faster in translation. No correlation is found between the translation methods preferred by post-editors and productivity. However, those that preferred working with Fuzzy matches are the least productive. When comparing with Fuzzy matches (of all categories including below 50 percent matches) with MT in those languages with best MT output (Chinese, Polish, German, Spanish, French and Italian), MT is more productive globally, and as productive as the matches in the 85-94 range in particular. The segment length that gives fastest results is 25 words for post-editing and 21 words for translation. A blind final translation quality evaluation was carried out for both types of translations and the results reveal that reviewers cannot tell the difference between post-edited text and human translation, and there is not a pattern that suggests loss of quality. Although the site does not provide all details of the experiment, we find the methodology and results interesting for the purpose of our project.

Apart from the papers reviewed in this section with the involvement of companies such as Adobe, Autodesk, Microsoft, SAP or Symantec, there have been a series of reports presented by private companies on the results obtained by integrating MT and post-editing in their localization workflows. This is the case of IBM (Roukos et al. 2011), Sybase (Bier and Herranz 2011), PayPal (Beregovaya and Yanishevsky 2010) and PayPal and Caterpillar (Dove et al. 2011), and CA (Paladini 2011), to name just a few recent presentations. The main issue with these reports is not, in our view, that companies might desire to report increases in productivity using MT in order to lower the prices, but the fact that often detailed data is not available (sometimes for

confidentiality issues) to fully explain the results, and that these measurements are taken in live settings with so many variables that conclusions on productivity due to the MT factor might be clouded by other variables such as fuzzy matches from TMs, translators' speed in straight translation, terminology research, type and order of files, number of participants in tests, or weak time measurements, to name just a few common characteristics. However, it is important to mention that these commercial reports are increasing and this shows the practical use of MT and post-editing in the "real world", the knowledge acquired in recent years in the commercial sector, and the inevitable fact that these measurements will be taken as standards if there is a lack of more scientific research, especially coming from Translation Studies. The most frequent languages used in these live experiments are German, Italian, French and Spanish but there is also data available in Chinese, Portuguese, Japanese and Russian, for example. The engines used are Moses, Lucy LT, or the company's own engine. One company may use different trained engines at the same time, depending on a particular language. The translation environment can be SDL Trados, IBM's TM2 environment or the company's own proprietary tools. The increases reported vary from 20 to 70 percent depending on language, domain, type of material (instructions, procedural documentation, user interface, and help systems), state of customization of the engine and subsequent quality of the raw output. The main conclusions to be drawn from these presentations are that MT contributes to the increase of productivity without a hindrance to quality if:

- Engines are trained with sufficient quality data;
- Engines are retrained periodically and cleaned up;
- Post-editors receive proper post-editing guidelines with specifications on quality expected and type of errors they will encounter;
- Initial time is invested in setting up the workflow and in sharing information about MT with all of those involved;
- Translation memories are well maintained and cleaned up regularly;
- The quality of the raw output is fully tested before implementation;
- The source text is improved through controlled language or other pre-editing techniques;
- Different languages are treated differently with different productivity expectations;
- Translators' accept MT;
- Translators are involved in the integration cycle;

- Payment is adequate according to carefully analyzed data and agreements with all those involved in the cycle.

It is also important to mention initiatives from private consultancy companies in the areas of post-editing, translation memories, and pricing structures. These studies are helping to bring quantitative data to support business decisions made in the localization industry. This is the case of TAUS (Translation Automation User Society), a community of users and providers of translation technologies and services. They are working towards good practices, quality assessment, identifying processes and exchanging experiences in machine translation and new translation technologies in general (TAUS 2012). Common Sense Advisory is a research and consulting firm with the goal of offering best practices and valuable insight about the localization process. They produce interesting reports on machine translation implementation and usage, post-editing and pricing systems (De Palma 2011, De Palma and Kelly 2009).

There are MT practitioners carrying out the important role of gathering data from all sources and distributing this area in the localization industry. This is the case of Asia Online through web seminars and newsletters (Asia Online 2012), and through blogs and knowledge-base sharing (Vashee 2012b) and Twitter.

2.2. Translation memories: Productivity, quality and processes

Despite the fact that translation memories have been in active use for the past two decades, there is a surprisingly low number of research projects on crucial areas such as productivity and quality of these intensively used systems (García 2008). We are including here the most relevant work on this area for our particular project, as well as studies on mental process and general use of TM technology that might be relevant for our data analysis because they incorporate translators' behaviors and opinions.

Gow (2003) investigated metrics for evaluating translation memory software. She compared two TM systems, SDL Trados and MultiTrans, using validity, reliability and efficient applicability as the evaluation criteria. She concludes that the sentence-based approach is better suited to text with many sentence repetitions and the Character String Based (CSB) approach is better for repetitions occurring within the sentence. She suggests that Trados could be better for experienced translators who are looking for an increase in speed and CBS for novice translators that need more terminological references. She favors the combination of both TM systems for optimal use.

Rieche (2004) in her Masters dissertation, studies the factors leading to quality problems in target texts when using translation memories. She analyzes segments from two systems - Trados Translator's Workbench 5.5 and Wordfast 4.0 - using the categorization: Terminology, Translation and Use of the language. The objective is not to find errors but to see if there are errors than can cause issues when leveraging and to set best practices to avoid error propagation. She finds a series of errors in the memories analyzed corresponding to the three categories (using both tools), independently of the size and content of those memories that can be propagated in future projects. She recommends, therefore, a systematic control of the translation memories through revision and maintenance, bearing in mind general principles rather than specific error classifications (such as LISA). She suggests that this should be carried out by experienced users once the project is completed. Although the study does not propose an automatic or systematic procedure to clean up memories, it draws the attention to the fact that a translation memory clean-up phase should be included in the localization process.

Dragsted (2004) carried out extensive empirical research as part of her PhD thesis on translation memories and segmentation. She compares the cognitive as well as temporal processing of a group of six professionals and six translation students. The results show that both groups, especially the professionals, perform differently when working with TMs and that they tend to change the sentence structure less frequently when working with the TM than when working on their own. She finds, with regards to productivity, that with standard texts professionals perform faster and the cognitive segmentation is different from that of students, but this difference seems to subside when work is done with difficult texts, or an unknown type of texts in the professional group. She points out that TM integration at the sentence level affects the translation process especially in professional translators, and she recommends that the retrieval of text should be done at the paragraph level.

Following up on Dragsted's work, Alves and Liparini Campos (2009) investigate the impact of translation memories and time pressure on types of external and internal support. They blend a series of classifications from different scholars for external and internal supports, together with pause classifications (in orientation and revision). They analyze the behavior of twelve translators with at least six years of professional experience, six from English to Brazilian Portuguese and six from German to Brazilian Portuguese, and they compare several environments: when translating on their own,

when translating with TMs, when translating under time pressure, and when translating with TMs and under time pressure. The translators worked with eight texts (four in English and four in German), approximately 4,000 words, using Translog and Translator's Workbench 7 (in combination with Camtasia, a software that allows recording of actions performed on the screen). The researchers carried out observations on-site and the translators commented on their translation process upon viewing the recordings. The results show that a separate orientation phase seldom occurs in the process of professional translators, and when it does occur, it tends to happen during the early drafting phase. During the drafting phase, orientation pauses are more frequent than in revision pauses, although there is extensive on-line revision. Revision occurs as a separate phase at the end of the translation process. Translators rely mostly on their own knowledge (internal support) to solve problems, with or without translation memories, in both the drafting and the revision phases. Translation memories do not appear to change the behavior of professional translators, despite the translators' relying more on the external support of the concordance feature, and internal support becomes more prevalent (to make decisions on the proposals made by the TMs). Time pressure has no effect on the type of support used but there are less revision pauses during drafting and revision, and more pressure on translators' to accept TM proposals without revising them. Alves and Liparini Campos conclude that translator training should take into account the importance of internal support "as the most productive type of support in all tasks combinations" (ibid: 209). Exploring how translators use technology and their own resources, both external and internal, is crucial to understand needs in future technology development and translation training. However, we feel that productivity and quality, for example, should be looked at in order to assess if the behavior of the translators is indeed "the most productive" and not just "the most frequent".

Bowker (2005) studied the correlation between translation productivity and errors when using translation memories. She carried out empirical research using nine translation students from French into English. She divided them into three groups: one did not use TMs, only reference material: the second used TMs: and the final group used TMs with seeded errors. The corpus was created using job advertisements for translation posts, as this was deemed of interest to translation students. The text volume was 387 words. In order to measure quality, Bowker used the ATIO (Association of Translators and Interpreters of Ontario) standards. Although she finds that using TMs increases productivity, she further discovers that the translators are not as critical of the

proposals made by the TMs, and they do not find most of the seeded errors. She points out that the continuous recycling of material in TMs using several translators affected the quality of TM databases, and therefore tighter quality control should be in place before attempting to use TM databases.

Wallis (2006) has studied the effect of translation methods on productivity, quality and translator satisfaction, using pre-translation or interactive mode in TM tools. He carried out a pilot project using four translation students from French into English in the domain of ultraviolet radiation and the ozone layer, using a TM tool called Fusion Translate (available at his university). He concludes that there are no differences in productivity when using the two methodologies (pre-translation and interactive), but a slight quality improvement and higher level of job satisfaction when using the interactive mode. The methodology used and the innovative approach are very relevant to our questions here and indeed to the localization industry. For our particular study, it is interesting to see that using the interactive mode available from the tool did not alter the productivity and only shows a slight quality improvement. It is also interesting that translators show more satisfaction when they are involved in the decision making process in the interactive option.

Vilanova (2006) in her Masters Dissertation studies the effects of translation memories on the target text at a linguistic level. She carried out a pilot experiment with three professional translators from English into Catalan. They translated the text they had previously translated using TMs, this time without a TM (1,000 words each). The texts are compared to establish if the texts done with a TM are closer at a linguistic and structural level to the source text than those translated without a TM. The results show that there is greater linguistic interference in the texts translated using TM than those translated from scratch. The translations done with the TMs contain more syntactical and lexical calques from the English, as well as calques in the macrostructure of the texts, while in their own translations, translators tend to follow the rules and structures of the target language. However, as Vilanova points out, the corpus is small, the texts are different and there are only three translators, therefore a larger sample is needed to draw final conclusions.

Ribas (2007) reports on an interesting experiment on the propagation of errors in TM systems. She compares a group of three students and three professionals and the data seemed to indicate that the TM system helps to propagate errors throughout translation projects. Although more conclusive data is required, this idea is of particular

interest to our study, since there is little research in the area of final quality in TM produced texts.

O'Brien et al. (2010) explore the usefulness of sub-segment matching through the analysis of how translators use the Concordance feature in TM technology. Six professional translators from German to English were recruited to translate a text of 424 words from the business domain, using SDL Trados 2007 with a memory containing 16 percent exact matches, 28 percent fuzzy matches and 56 percent generating concordance matches. One group translated with the Concordance feature enabled and the other group with the feature disabled. Both groups were recorded using a Tobii 1750 eye tracker and they answered a post-task questionnaire. The final translations were analyzed with the LISA QA Model. The results show that translators use the Concordance feature as a sub-segment matcher. Furthermore, the data suggest that when the Concordance feature is enabled, translators use the information provided even more than the full TM proposals. Translators with the Concordance feature enabled take longer than those with the feature disabled but the final quality is higher. Translators think that this feature is useful for terminology and context but not necessarily to enhance their productivity. Therefore, they do not agree with the possibility of offering discounts over sub-segments matches (something that the authors suggest might have been in store with the new version of SDL Trados, 2009). As the researchers mention, this type of research is crucial not only for TM users and developers but also for the MT community, as both systems are progressively used in conjunction in the localization industry. It is true that the number of participants is low but the design and the findings are very relevant.

Yamada (2011a) investigates how productivity is affected by different kinds of TMs. He carries out a pilot study with eight Master translation students with varying professional backgrounds as well as working languages (four were native Japanese speakers and four were English native speakers). Participants are provided with a training session and “exercise lessons” (ibid: 65) on the SDL Trados 2007 before translating the same source text (500 words) with two translation memories: one TM is referred to as free-translation content and contains segments from real localization project, the other TM is referred as literal translation content and in this case the “freer” renditions of the source target in the original TM are eliminated from the target text to create a more “literal” content. All participants’ actions on screen are recorded using BBFlashback. The results show that the difference in average speed between the two

TMs is not highly significant (1:04:22 with the free content and 1:05.44 with the literal content). This is partially due to the fact that the fastest translator is in the free content group and the slowest in the literal content group. Although, this is expected when measuring productivity with such a small sample of participants (four in each group), Yamada's analysis of each participant's performance per fuzzy match is more revealing. It shows that translators in the free content group experience a flat productivity even with different degrees of fuzzy matches (the theory would be that the higher the fuzzy match, the higher the productivity) when compare to the literal content group thus suggestion that the freer content "may have reduced the translator's segmentation recognition speed in higher match categories" (ibid: 71). Although Yamada concludes that the production speed in fuzzy matches is faster using the more literal content, the fastest translator (in the freer group) is not included, thus reducing even further the sample. He also suggests that "if free segments are put into the TM database, there is a chance that this may adversely affect the translator's performance" (ibid: 72). This might indicate a need for a "controlled" target text, a series of instructions or detailed style guidelines given to translators in order to translate more literally. We are not referring here to introducing errors in the translation as a result of mistranslations but to avoiding more "creative" renderings of the source text that might have an equivalent meaning. This would increase productivity when leveraging TMs, but also, in the future, to train MT engines more effectively.

Yamada (2011b) further explores revision when integrating TM and MT in the translation process as part of his doctoral thesis. He studies five professional translators and 18 students in the English to Japanese language pair (numbers include previous pilot). The participants worked on SDL Trados2007, Google Translate and they are recorded (keyboard and mouse movements as well as time stamps) using BB Flashback. He uses GTM (General Text Matcher) (Turian et al. 2003, Melamed et al. 2003) to measure the "revision amount between two texts" (ibid: ii). The results show that professional translators are faster than students when translating, and both groups show similar productivity increases when using TMs. The amount of time spent on the revision phase of the translation is higher for students than for professionals, 44 and 24 percent respectively, but professional spend more time in the drafting phase. When using TMs the revision time decreases in both groups while time increases during drafting. The GTM scores appear to correlate with the fuzzy match ranges, that is, the higher the score (higher similarity) the higher the fuzzy match. The GTM scores are

higher for professionals than for students, indicating that the professionals make fewer changes. He establishes that if GTM is “over 0.464, it may be faster to revise the fuzzy matches, and if it is below 0.464, it is easier to translate the source text from scratch.” (ibid: 146). Yamada includes MT as a second phase of this project with 13 participants (eight students and five professionals) in two different experiments. The professionals received an assignment brief informing them that there would be MT segments. The results show no productivity gain when using MT in the professional group (the student group is not measured), and low GTM scores indicating that translators make a substantial number of edits to the output. He concludes that translator would prefer using MT if the GTM score is 0.464 or higher, and this would be similar to a 55 to 60 percent fuzzy match. The study gives relevant information when using TMs and GTM scores correlations, however we feel the methodology used and the series of experiments are combined in a way that they generate many variables difficult to control, compare and correlate. For example, TM increases the productivity but this is somewhat logical if 100 percent matches are included. Google Translate is used as the SMT engine but no score of the quality is used. This is not a customized engine used in professional environment so the conclusions drawn on the correlation of GTM and productivity cannot be applied to a professional environment. Also, the volumes used are rather small, especially for the post-editing phase. Finally, Yamada himself measures quality of the post-edited material without a clear explanation of the criteria used.

The research group TRACE (Spanish acronym for Computer Assisted Translation (CAT), Quality and Evaluation) from the Universitat Autònoma de Barcelona studies the influence of CAT tools on translation processes. The group has published different papers considering different aspects of this influence. Torres-Hostench et al. (2010) present results on the study of explicitations, interference and textuality phenomena in relation to the use of CAT tools (from English to Spanish). The results given in this paper are from a pilot project carried out with eighteen MA translation students. The students are asked to translate three different texts using three possible scenarios (MS Word, SDL Trados 2007 and SDL TagEditor, without a translation memory). A screen recording tool and a keyboard-logging tool are used in each computer. The results show that there are no clear differences in explicitation when using any of the three environments, there are partial differences with regards to interference and finally there are differences in textually. When using a TM

environment, translators are more faithful to the source text punctuation rules than to the target text. Martín-Mor (2011) presents the specific results related to the linguistic interference as part of his doctoral thesis, using both novice and professional translators (freelance and in-house translators). He finds that the translator's profile, the tool and the position (the order in which the task is done) significantly affect linguistic interference. Interestingly, the CAT tools can cause significantly more interferences (especially with regard to spelling, typography and cohesion) in novice translators, since professionals tend to compensate for this effect. On the other hand, the translators' attention to the segment might cause less interference on other aspects because it requires them "to compulsory perform an action" (ibid: 405) before going to the next segment. In-house translators show more lexical interference than the other two groups (freelance and novice) possibly because they work consistently on texts of a highly technical language. Regarding the position, CAT tools have an effect on subsequent translations in non-CAT tools, and vice versa, but the actual type of tool (SDL Trados or TagEditor) have a similar distribution of interferences. Finally, Mesa (2011) describes the methodological results provided by a pilot study that measures the impact of CAT tools on the use of explicitations. Although the methodology explained is thorough, no quantitative or qualitative results are given in this particular paper.

Teixeira (2011) explores the effects of provenance knowledge on translation performance in an integrated TM/MT environment. His aim is to find out if knowing the origin and level of fuzzy of a segment (TM or MT) has an impact on the speed and quality. He carried out a pilot test with two translators from English into Spanish translating 500 words in two different environments (with and without provenance information). He measured the time through screen recording and key logging using BB Flashback. He concludes that the overall speed is not significantly different when using the two environments and the quality is of comparable level. The exact matches, however, are processed faster when the provenance is known. However, as he points out, this is only a small pilot with two participants and two reviewers, and the results are inconclusive.

Moorkens (2011) explores the assumption of consistency in TMs. With this aim in mind, he analyzes four sets of TM data (two sets of English into Japanese and two sets of English into German) according to inconsistencies in both source and target, inconsistencies in source but not in target, inconsistencies in target and not in source and consistent source and target. The results show that there are inconsistencies across

all memory sets in all categories. He finds inconsistencies in nouns, inconsistent use of spaces and punctuation that have a negative effect in leverage even if not in the actual translation, explicitations that might render the text not useful for leverage as an item might be added that makes that sentence/segment clearer in that particular instance but that might not be appropriate in others, insertion of comments. All of these inconsistencies, he concludes will have a negative effect on the final text, especially in those segments where translators will not be permitted to make changes (for example 100 percent matches). We believe it is of utmost importance to explore how translation memories impact the quality of the final text or the way in which translators process the leveraged text. However, we are also aware that a lot of inconsistencies are caused by changing terminological or stylistic instructions over time by the customers themselves and we do not think this was considered in this study. On the other hand, it is also interesting to question if certain type of inconsistencies should be “permitted” and more attention be paid to the actual readability of the text. Having said this, a consistent and cleaned memory plays a fundamental role in the performance of MT engines and a better understanding of the type of inconsistencies to systematically avoid them is necessary.

Christensen and Schjoldager (2011) carried out a preliminary pilot project with twenty-two MA translation students to find out their experience after an introductory course on TM translation by means of an on-line questionnaire. There was a practical assignment using SDL Trados Translator’s Workbench (2007) that consisted of a translation from English to Danish of instructions for a mobile phone using a memory populated with very few segments. Afterwards, students were asked to fill in an on-line questionnaire. The results show that students find working with TMs different than working on their own, and that they see positive aspects (mainly speed, organization, efficiency and consistency) and negative aspects (less thinking on their own, potential danger of accepting the wrong translation). As the authors say “many regretted a general loss of control” (ibid: 124). Regarding sentence segmentation, students are more unaware of the effects, some think it is an advantage because it is a useful way to classify the task but they also comment that it might force them not to look at the contextual and functional aspects of the translation. The authors suggest that the greatest impact of TM occurs during the drafting phase of the translation process because translators might accept proposals without questioning them and also that segmentation imposes a way of dealing with the text to the students that might be different from their

own. They comment also that the comprehension phase is less thorough and that less micro-strategic decisions are required in the transfer phase because of the leverage from other translations. The results seem to imply that the planning phase is non-existent and that students suggest that a post-drafting phase should change to further review the target text. Although, these findings are opinions that students gave and not actual empirical findings, and also we are not presented with a comparison to see how translators deal with a text without TMs (perhaps certain phases of the translation process are also different in professional environment even without TMs) we find the questionnaire interesting in order to understand TM technology impact on translators as this might also be potentially applicable to MT. If the segments generated are of a very high standard, translators might accept blindly proposals causing similar type of changes in the translation process and they might also regret this loss of control of the translation process in general.

Chapter 3: Methodological considerations

In this chapter, we will describe the background to the research project, the hypotheses formulated and their operationalization, the methodology used for the project, how the methodology was tested, the validity and generalizability of the research, the threads of this validity, and a brief description of the project development.

3.1. Pilot Project

A pilot project, the details of which are described in Guerberof 2008, was carried out with eight subjects. The results showed that post-editing MT segments was in fact faster on average (mean value considered) than post-editing TM segments. The mean value was 13.86 words per minute with MT as opposed to 12.14 words per minute with TM (11.87 words for New segments). Nonetheless, the data dispersion was very high among participants, with high standard deviations and great differences between maximum and minimum values, suggesting high subject dependency. Still we observed that translators with less experience and lower processing speeds were likely to have similar processing speeds when using both aids, MT and TM. On the other hand, translators with more experience had higher processing speeds when using MT. Further, some of the fastest post-editors did not appear to benefit from any translation aid as they seemed to process the segments faster without aids. The productivity gain on average was between 13 and 25 percent for MT segments and between 10 and 18 percent for TM segments.

The final quality of the samples and the results obtained on errors were quite striking. The number of errors in TM segments was higher than in New (by 141 percent) and than in MT segments (by 91 percent). On the other hand, the number of errors in New and MT segments was quite close, despite being slightly higher in MT. Another important point was that the number of Accuracy errors was higher than any of the other type of errors (44 percent), particularly in TM segments. This led us to believe that translators accepted more readily proposals from the TM without necessarily questioning the content, because the natural flow of the sentences was similar to a human translation, while the errors in MT were obvious and easier to detect. Translators also showed more terminological errors in MT and TM than in New segments. This suggested that translators accepted the proposals made without necessarily checking the glossaries. Mistranslation errors were lower in MT than in New and TM segments, and

this might indicate that MT helps to clarify difficult aspects of the source text. More data, however, was needed to test this trend further. Post-editors with higher total processing speeds had fewer errors in the samples and this showed that spending more time on revision did not necessarily improve translation quality.

Translators' experience had an impact on the processing speed. Translators with experience performed faster on average. If we looked at the number of years of experience in localization, domain, tools and post-editing MT output, we observed an increasing curve up to the 5-10 year range and then a drop in the speed. The number of errors was higher in experienced translators by a very small margin, and there were more errors in MT segments. This might have indicated that experienced translators grow accustomed to errors in MT output. On the other hand, translators with less experience had more errors in New segments than in MT segments, which seemed to indicate that MT has a leveling effect on their quality.

We felt, however, that the sample of eight participants was a highly limiting factor. It was necessary to explore further the relationship between productivity, quality and experience with a greater number of participants and with more qualitative data from questionnaires and debriefings.

3.2. Hypotheses

As a result of the pilot study (Guerberof 2008) and our practical experience in localization, we have formulated the following hypotheses. These can be grouped under the general concepts of Productivity, Quality and Experience.

These hypotheses will be tested for the English-to-Spanish language pair, supply chain management content and a Moses statistical machine translation engine trained with more than 1.9 million words of bilingual corpora and with a high quality output (measured with BLEU score and human evaluation as seen in section 3.4.3.4).

3.2.1. Hypothesis 1: Productivity

The time invested in post-editing machine-translated text will correspond to the time invested in editing fuzzy-matched text corresponding to the 85-94 percent range.

Here productivity will be measured as words per minute. In other words, MT output provides levels of usefulness similar to those of TM fuzzy matches in the region of 85-94 percent (measured according to the SDL Trados matching algorithm). Not only

would it require less effort, measured in time, to translate MT segments than to translate No match segments but this would be an effort equivalent to the translation of Fuzzy match segments (in the 85-94 percent range).

3.2.1.1. Sub-hypothesis

The translators with higher processing speeds, in words per minute, when translating the “No match” segments will show less productivity gain when post-editing the proposed text from MT or TM than the translators with lower processing speeds when working with the same set of segments.

As explained in above, in our pilot study, we observed that translators with higher processing speeds when translating No match segments had less productivity gain when using MT and TM segments than those that had lower processing speeds, hence the need to test this sub-hypothesis in a larger scale project.

3.2.2. Hypothesis: Quality

To make a more comprehensive measurement of these translation processes, it is necessary to consider the final quality. If the time necessary to review MT segments is shorter than the time necessary to review No match or Fuzzy match segments, but there are more errors in the final target text, the productivity gain made during the post-editing phase would be reduced or even negated. Therefore, our second hypothesis is that *the final quality of the revised target segments translated using MT technology is higher, if measured in number of errors, than the final quality of revised Fuzzy match segments and lower than the final quality of revised No match segments.*

The quality will be measured according to the number of errors in the final texts: the higher the number of errors, the lower the quality, and vice versa.

3.2.2.1. Sub-hypothesis

Translators, with higher overall processing speeds when using MT or TM technology, will have fewer errors than those with lower processing speeds.

As explained above, in our pilot study, we observed that translators with higher processing speeds had fewer errors in the final target text than those with lower processing speeds. Now, we want to test this trend with more translators.

3.2.3. Hypothesis 3: Experience

Localization has a strong technical component because of the nature of the content translated as well as the tools required. On many occasions this experience is associated with speed, that is, the more experience in localization, tools used and domain, the less time will be needed to complete a project. Therefore, our third hypothesis proposes that *the greater the experience of the translator, the greater the productivity in post-editing MT match and Fuzzy match segments.*

3.2.3.1. Sub-hypothesis

This technical experience will not have an impact on the quality (measured in number of errors).

As we mentioned in the previous two sub-hypothesis, this pattern was observed in the pilot project where translators with more experience measured in years processed the segments faster but they did not necessarily produce a final target text with fewer errors than those with less experience.

3.3. Variables and operationalization

In order to operationalize these hypotheses we will use time spent to measure the processing effort in three different segment categories: No match, MT match and Fuzzy match segments. In order to measure time without other variables that might distort the result, we will use a similar amount of words for each category and the same type text with one group of professional translators (we will explain our criteria for selecting the professional translators in section 3.4.3.1). We will measure the time in words per minute to reflect the time employed to process the segments. The amount of words will not cover a full day's work (according to the standard measurement in localization) and therefore we will not use words per day, as results might be distorted. We will, however, present an extrapolation of speed per day for orientative purposes.

Quality will be measured according to the number of errors in the target text in three different segment categories: No match, MT match and Fuzzy match), to determine if a higher or a lower productivity has an impact on quality. The errors will be defined according to the LISA QA Model (see The LISA QA Model in Appendix A). We will use a form slightly modified for the purpose of this study so that reviewers can distinguish between No match, Fuzzy match and MT match.

Experience will be measured according to years of experience in localization, subject matter, tools, and in post-editing. We will also consider the tasks (translating, reviewing, post-editing, terminological tasks) translators perform frequently on the job, estimation of daily throughputs and average typing speed. We will obtain the information through a post-assignment questionnaire/survey. The degree of experience will be matched against processing speed and number of errors to determine if more or less experience in these areas corresponds to a higher or lower processing speed and number of errors.

3.4. Methodology

3.4.1. Mixed methods approach

In our pilot project (section 3.1) we focused on a quantitative analysis to test our hypotheses. However, there was not sufficient qualitative data to explain some of the phenomena observed when examining the quantitative results. The survey used to gather information on experience could potentially be expanded to gather more data on the participants' working styles, and their opinions about machine translation and the profession, debriefings could give us more details about the participants' experience during the assignment.

For this reason, we decided to use a mixed methods approach in this project consisting of two phases: quantitative followed by qualitative. The first phase will gather quantitative data that will be analyzed by means of descriptive and inferential statistics; the second part, which will occur immediately after we finish collecting the quantitative data, will gather qualitative results by means of a survey and debriefings with the translators and reviewers, which will help to explain the results obtained in the quantitative analysis. We will thus use a simultaneous or concurrent design (Creswell 2003, Creswell and Plano Clark 2007). Due to the nature of our exploration (see section 3.2), the quantitative data will have priority over the qualitative data, in a relationship that might be expressed symbolically as: QUAN + qual (Morse 2003, Morse and Niehaus 2009). The data will be combined during the interpretation or final discussion phase.

Since the qualitative data will not be fully integrated, the quantitative data will not be analyzed in such a way that we would then collect the qualitative data depending

on the results found; the qualitative data will serve to explain and gather information of the assignment and participants' general opinions. Tashakkori and Teddlie (2009) call this type of mixed research method a "Quasi-mixed design" since "two types of data are collected (QUAN, QUAL), with little or no integration of the two types of findings" (ibid: 142). There are two types of data in our study, QUAN is the dominant or driver, and QUAL is the explanatory or complementary.

3.4.2. Overview of experimental project

We would like to offer a brief roadmap of the empirical project as a way of better understanding each of the different subsections that will be explained in detailed below.

The experiment was performed with 24 translators and three reviewers. One participant carried out the preliminary test and three reviewers revised the work done by this one translator so the clarity of the instructions could be assessed and necessary changes implemented in the final project. The post-editors used a web-based post-editing tool designed by CrossLang to post-edit and translate a text from English into Spanish. The text had 2,124 words: 749 words of No match segments (new text to translate), 618 words of translation memory segments (SDL Trados 2007 was used to create the fuzzy matches) and 757 words of machine-translated segments (a trained Moses statistical-base engine was used to create the output). We selected a supply-chain software product for the corpus, as we wanted to use authentic content from the localization industry. At the end of their assignment, the participants filled in an on-line questionnaire with information related to the experiment and their own experience in the field. The final output was then revised by the three reviewers, who counted the errors using the LISA QA model criteria, and filled in an on-line questionnaire. Twenty-three participants were debriefed after the assignment to gather their opinions and impressions on the assignment. All data were processed together with a team of statisticians. Final results were analyzed and conclusions drawn.

3.4.3. Data for quantitative analysis

This section includes the data used for the quantitative and analysis although part of this data was also used in the qualitative analysis.

3.4.3.1. Sample

For this project we had a group of 25 professional translators and six professional reviewers. One of the translators and three reviewers participated in the testing phase only; the remaining 24 translators and three reviewers participated in the full project. They were situated in different locations and time zones. They were contacted by email at all times and they received no specific training, only a single set of instructions (see Appendix C). Since the research project involved professional translators, the Vendor Management team at a professional language service provider (HiSoft Spain) cooperated to liaise with translators and reviewers. The team followed the criteria given to them, as if this research project was “the production project”, and the research team was “the project managers”.

3.4.3.1.1. Criteria for selecting translators

The professional translators from English to Spanish were selected from those approved in the HiSoft database. From this population we eliminated the translators that had been part of the MicroStrategy localization team at HiSoft or that had participated in the pilot research project (Guerberof 2008), in order to avoid an increase in productivity due to previous knowledge of the domain or the post-editing tool. Any such increase could affect the validity of the variables we were interested in exploring. Since post-editing is a relatively new task for language service providers and the number of post-editors available is limited, the Vendor Management team normally contacts translators with or without experience when a new post-editing project arrives. Further, we wanted to see the impact experience had on productivity and quality, so it was an advantage to have a sample with a varied background. Therefore, we followed exactly the same criteria: no post-editing experience was required a priori but translators with post-editing experience were not to be discarded.

3.4.3.1.2. Criteria for selecting reviewers

The professional translators and reviewers from English to Spanish had to have at least three years' experience in localization (software, help and/or documentation) and in Computer Aided Translation Tools (SDL Trados, Déjà Vu, MemoQ or similar tools). Familiarity with tools is an indication of familiarity with translation memories and post-editing, and reviewers needed to be aware of the environment translators were working on to assess the texts produced by them. The reviewers should also have at least six months' working experience in MT post-editing and in Business Intelligence software

translation. These criteria were applied so reviewers were sufficiently familiar with the task in hand as to evaluate the work done by the translators without introducing unnecessary changes and at the same time with sufficient technical knowledge to perform the task at the standard review speed.

3.4.3.1.3. Selection process

The Vendor Management team contacted the translators that had been working for HiSoft since 2009 in the language combination English to Spanish and that had not worked on MicroStrategy projects. At this point 80 translators were found. The VM team eliminated those that had participated in the pilot project, those without long-term experience in working for HiSoft, and those that had produced poor quality (according to their internal review methodology based on the LISA metric) in previous projects. After some of these translators were discarded, the VM team sent an email to the remaining 40 translators. Some confirmed their availability and others did not, thus reducing the list initially contacted. The VM team also contacted ten additional translators that had either passed the translation test or that had worked for HiSoft as reviewers. The translators and reviewers available (until the figure of 31 was reached) were informed about the nature of the project (see Appendix B), the rate (the fee agreed was the standard full rate per word for this language combination at HiSoft) and the time frame, as in a standard localization project. If they agreed to participate they were asked to sign a Research Participant Release Form where it was clearly stated that their participation was voluntary and where they granted permission for the evaluation of the data without identifiable information in regard to their name.

3.4.3.2. Corpus

MicroStrategy, a company specialized in business intelligence technology, that provides integrated reporting, analysis, and monitoring software, was contacted in order to obtain permission to use their bilingual corpus for the language pair English-into-Spanish. We decided to contact MicroStrategy because the type of content produced was frequently translated in the localization industry and, more importantly, we knew that the volume of bilingual text available in this language was sufficient to train a statistics based machine translation (the MT provider had indicated that in order to achieve an acceptable MT output quality, more than 500,000 words of bilingual corpora was needed). Furthermore, although it is generally difficult to have access to proprietary

material from software developers, in this case, the legal department at MicroStrategy granted permission to work with their content.

3.4.3.3. Statistical Machine Translation (SMT) engine

CrossLang, a provider of language automation solutions, was contacted to participate in the project. They trained a statistical engine and provided a web based on-line post-editing tool. CrossLang has access to several statistical engines but we decided to start training Moses (Koehn et al. 2007) because it is an OpenSource engine that is increasingly being used by Language Service Providers as opposed to proprietary engines that would require a substantial initial financial investment (TAUS 2012). Plitt and Masselot (2010) also chose Moses for their experiment because it “is easily expandable across several languages at once”, “we possess considerable amount of high quality legacy translations” and “it would have been difficult to reach return on investment with a commercial machine translation system” (ibid: 8). If Moses, after being trained with the MicroStrategy data, had not given appropriate results (after the evaluation explained in section 3.4.3.4) we would have considered using another engine, possibly a proprietary engine.

Moses is a statistical machine translation system that allows automatic training of translation models for any language pair. All that is required is a collection of translated texts (parallel corpus). Then, a search algorithm quickly looks for the highest probability translation among an exponential number of choices (Koehn 2012a).

3.4.3.3.1. Training the engine

In order to train the engine, we used a translation memory (TM) in SDL Trados 2007, and three glossaries (two in Excel and one in xml format). The memory had 173,255 segments and approximately 1,970,800 words (English source) and it contained multiple translations. This meant that for one source segment, the translation memory might have contained several target segments. In this case, some English titles could either remain in English or be translated, depending on the given product and context, thus creating several target segments. The Excel glossaries contained 610 entries and 94 entries of core terminology; the xml file contained 9,106 entries (software strings in xml format).

Table 1 shows the components in the translation memory.

Component	Products
Guides	Basic Reporting Guide
Guides	Advanced Reporting Guide
Guides	Documentation Creation Guide
GUI Strings	GUI Strings
Online Help	Online Help
Guides	Mobile Guide
Training Material	Web Online Courses
Training Material	Instructor Lead Courses
Guides	iPad Guide

Table 1: Translation memory components

3.4.3.4. BLEU score and human evaluation

BLEU is an automatic evaluation metric developed to help researchers test the quality of MT output faster than with a human evaluation process. The metric is based on the idea that “the closer a machine translation is to a professional human translation, the better it is” (Papineni et al. 2002: 311). The score is calculated using human references as a baseline and then it gives a value ranging from 0 to 1: the closer the values are to 1 the better the MT output is considered to be.

The BLEU score for our Moses project was 0.6. This is calculated by reserving 1,000 randomly selected sentences of the data available for building the system (translation memories) and then using them as a “test set”. As we saw above, the closer the value is to 1, the better the MT output is considered to be. A score of 1 will mean that no changes are necessary as the quality would be a “perfect” human rendering. BLEU scores produce a number but this number does not mean that it is a percentage of accuracy – it is just an indication of how close it is to a professional human translation. It is important to consider that “even two competent human translations of the exact same material may only score in the 0.6 or 0.7 if they use different vocabulary and phrasing” (Vashee 2012a). In order to further confirm these results, we performed a human quality test using an MT Evaluation Tool designed by CrossLang. Figure 1 shows a sample of this tool.

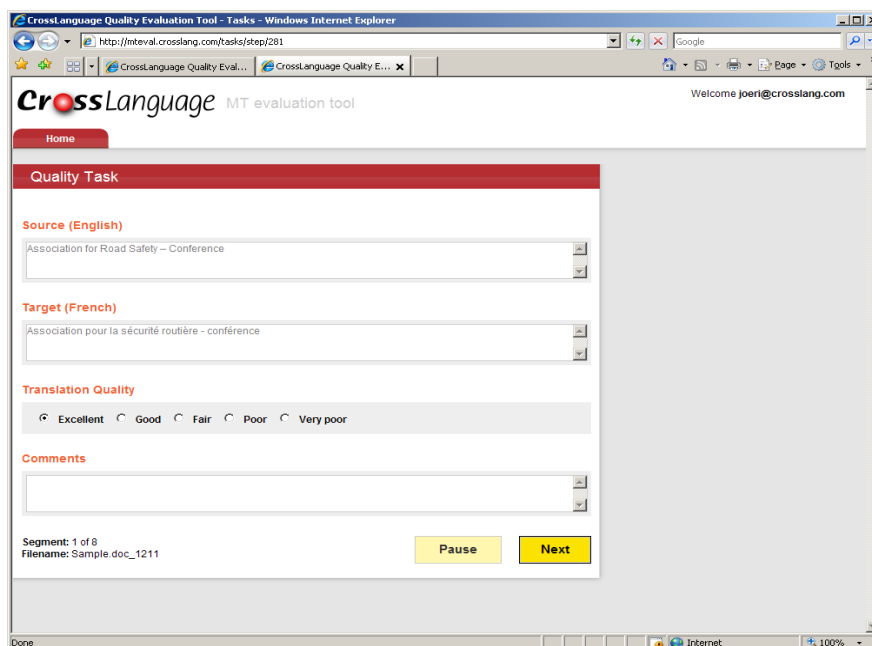


Figure 1: Screen-shot of the MT evaluation tool used

We tested 200 segments out of a sample of 900 segments and we categorized each segment as Excellent, Good, Fair and Very Poor quality, providing comments and justifications. Table 2 provides a description for each value.

Values	Description
Excellent (5)	Read the MT output first. Then read the source text (ST). All meaning expressed in source fragment appears in the translation fragment. Your understanding is not improved by reading the ST because the MT output is satisfactory and would not need to be modified (grammatically correct/proper terminology is used/maybe not stylistically perfect but fulfills the main objective, i.e. transferring accurately all information).
Good (4)	Read the MT output first. Then read the source text. Most meaning expressed in source fragment appears in the translation fragment. Your understanding is not improved by reading the ST even though the MT output contains minor grammatical mistakes (word order/punctuation errors/word formation/morphology). You would not need to refer to the ST to correct these mistakes.
Fair (3)	Read the MT output first. Then read the source text. Much meaning expressed in source fragment appears in the translation fragment. However, your understanding is improved by reading the ST allowing you to correct minor grammatical mistakes in the MT output (word order/punctuation errors/word formation/morphology). You would need to refer to the ST to correct these mistakes.
Poor (2)	Read the MT output first. Then read the source text. Little meaning expressed in source fragment appears in the translation fragment. Your understanding is improved considerably by reading the ST, due to significant errors in the MT output (textual and syntactical coherence/textual pragmatics/word formation/morphology). You would have to re-read the ST a few times to correct these errors in the MT output.

Very poor (1)	Read the MT output first. Then read the source text. None of the meaning expressed in source fragment appears in the translation fragment. Your understanding only derives from reading the ST, as you could not understand the MT output. It contained serious errors in any of the categories listed above, including wrong POS. You could only produce a translation by dismissing most of the MT output and/or re-translating from scratch.
------------------	---

Table 2: Human evaluation criteria

The output had a mean score of 4.5 out of 5 which meant that the quality was between the values Good and Excellent. Figure 2 shows a breakdown of the human evaluation results.

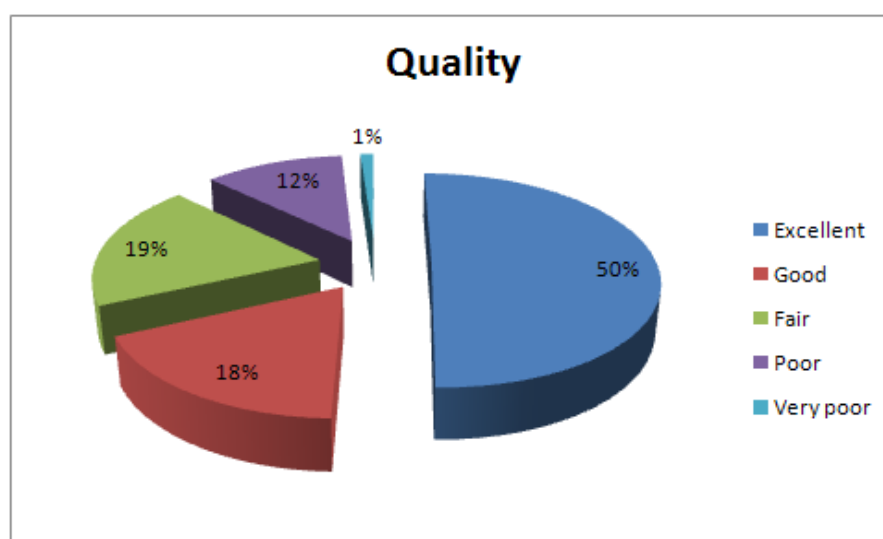


Figure 2: Human evaluation of MT output

Having the data from the BLEU score and the human evaluation, Moses was selected as the engine to use for the project.

3.4.3.5. Dataset

Once the Moses engine was trained with the previous translation memory and glossaries, we used a new set of data to create the fuzzy matches and the machine-translated segments for this particular project. The file set was a help system and user interface strings that came from MicroStrategy as part of a standard translation project. The name of the project was: Online Help 9.1 and Strings 9.1, batch 1 and batch 2. We will refer to it as “project 9.1”. The penalties used when pre-translating were: 1 percent penalty for missing formatting, 1 percent penalty for incorrect formatting, 1 percent for multiple translations. Table 3 and Table 4 show the word-counts obtained with SDL Trados 7.0.

Analyze Total	(652 files):		
Match Types	Segments	Words	Percent
Context TM	0	0	0
Repetitions	1,058	8,490	2
100%	19,812	216,730	63
95% - 99%	6,202	71,903	21
85% - 94%	393	5,487	2
75% - 84%	348	4,419	1
50% - 74%	122	1,530	0
No Match	2,199	33,661	11
Total	30,134	342,220	100
Chars/Word	4.87		
Chars Total	1,669,342		

Table 3: Analysis of online help for project 9.1

In this case, the on-line help system had 652 files, and although the total number of words was 342,220 words, only 33,661 of those were new words. The rest of the source words were already found in the memory in different types of fuzzy matches (100, 95-99 percent, etc.). Repetitions refer to the segments that are repeated within the No match category (so this category would contain the second segment repeated). Context TM refers to those 100 percent matches that are also in the same context as they were in the previous version (same segment before and after).

Analyze Total	
Match Types	Words
Context TM	0
Repetitions	88
100%	295
95% - 99%	207
85% - 94%	302
75% - 84%	221
50% - 74%	63
No Match	2,362
Total	3,538

Table 4: Analysis of software strings for project 9.1

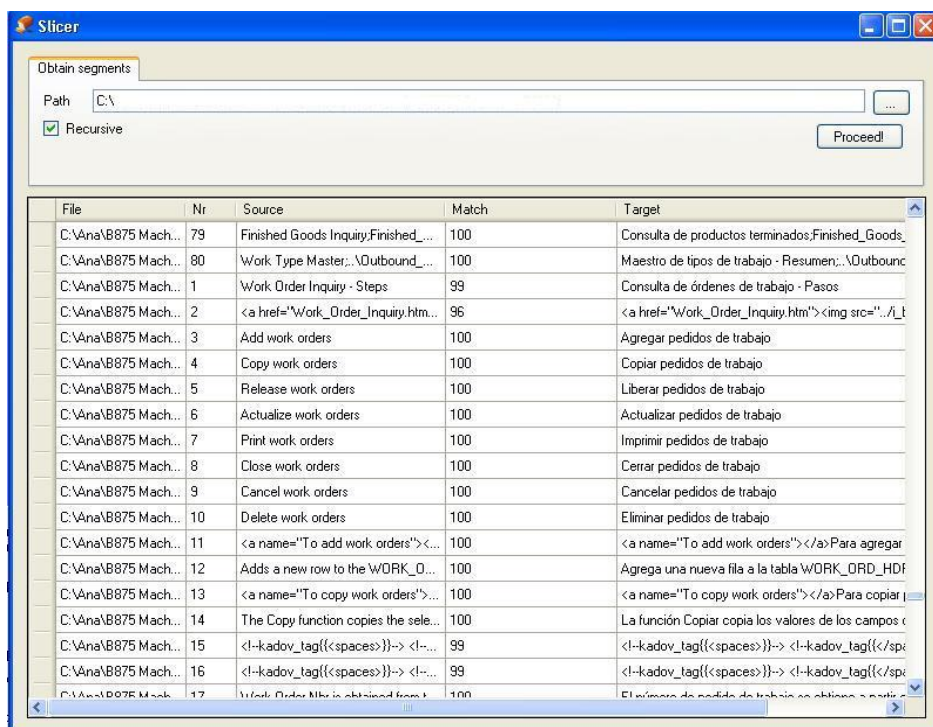
In this case, there was one file to analyze: the total number of words was 3,538, 2,362 of which were new. Since there were fewer words and fuzzy matches in this component, most of the strings used in this project were taken from the on-line help system. CrossLang uploaded these segments into the post-editing tool. Our main interest was to compare the MT segments with TM segments in the 85 to 94 percent range.

Therefore we had to prepare the corpus in order to be able to upload only these types of segments into the post-editing tool.

3.4.3.5.1. Selecting the fuzzy matches

These Fuzzy match segments were coming from the translation memory and had thus been translated by professional translators who had uploaded their translations in previous projects using SDL Trados.

For our research, we needed to create a file containing segments in the 85-94 percent category to feed these fuzzy matches into the tool. To prepare the file, we pre-translated the new files (project 9.1 as explained in section 3.4.3.5) with a previous memory in order to obtain fuzzy matches using the option Pre-translate in Trados. We exported all segment pairs together with their corresponding fuzzy level (54, 75, 86 and so on) to Excel. This was done with a small tool called Slicer specifically created for this purpose. This tool was created with .net in order to extract the bilingual pairs from the ttx files, together with their level of concordance, and to place them into a table, as seen in Figure 3. From Slicer we pasted all the segments into Excel, and then sorted the segments according to their level of fuzzy matches.



The screenshot shows the Slicer application window. At the top, there is a section titled "Obtain segments" with a "Path" field set to "C:\\" and a "Recursive" checkbox checked. A "Proceed!" button is located to the right. Below this is a table with the following columns: File, Nr, Source, Match, and Target. The table contains 17 rows of data, each representing a segment pair with its corresponding fuzzy match percentage.

File	Nr	Source	Match	Target		
C:\Ana\B875 Mach...	79	Finished Goods Inquiry;Finished_...	100	Consulta de productos terminados;Finished_Goods_...		
C:\Ana\B875 Mach...	80	Work Type Master;.\Outbound_...	100	Maestro de tipos de trabajo - Resumen;.\Outbound...		
C:\Ana\B875 Mach...	1	Work Order Inquiry - Steps	99	Consulta de órdenes de trabajo - Pasos		
C:\Ana\B875 Mach...	2	<a href="Work_Order_Inquiry.htm...	96	<...	100	Para agregar
C:\Ana\B875 Mach...	12	Adds a new row to the WORK_O...	100	Agrega una nueva fila a la tabla WORK_ORD_HDF		
C:\Ana\B875 Mach...	13	...	100	Para copiar		
C:\Ana\B875 Mach...	14	The Copy function copies the sele...	100	La función Copiar copia los valores de los campos		
C:\Ana\B875 Mach...	15	<!--kadv_tag({<spaces})--> <!--...	99	<!--kadv_tag({<spaces})--> <!--kadv_tag({</sp...		
C:\Ana\B875 Mach...	16	<!--kadv_tag({<spaces})--> <!--...	99	<!--kadv_tag({<spaces})--> <!--kadv_tag({</sp...		
C:\Ana\B875 Mach...	17	Work Order Mtr is obtained from	100	El número de pedido de trabajo se obtiene a partir		

Figure 3: Segments exported in Slicer

We took the 0-50 percent range to use with the MT engine and the 85-94 range to create the fuzzy matches. We deleted all duplicates and all segments with length below 4 and above 26 words. The segments of less than four words and more than 25 words were eliminated. The three-word segments, without context, can generate a lot of doubts during translation, whereas in the case of the segments over 26 words, there are very few in the corpus and we did not have sufficient material to replicate them in all categories (No, Fuzzy and MT match). We replaced all tags with place holders ({ph}) and then we applied the corpus sampler (a macro designed by CrossLang). The sampler creates a histogram with lengths of the corpus and it applies the same length distribution to the sample. We applied this same pattern to each category (No match, Fuzzy match and MT match) so that the distribution in each segment was balanced. Figure 4 shows this distribution in the histogram for the No match category in the sample.

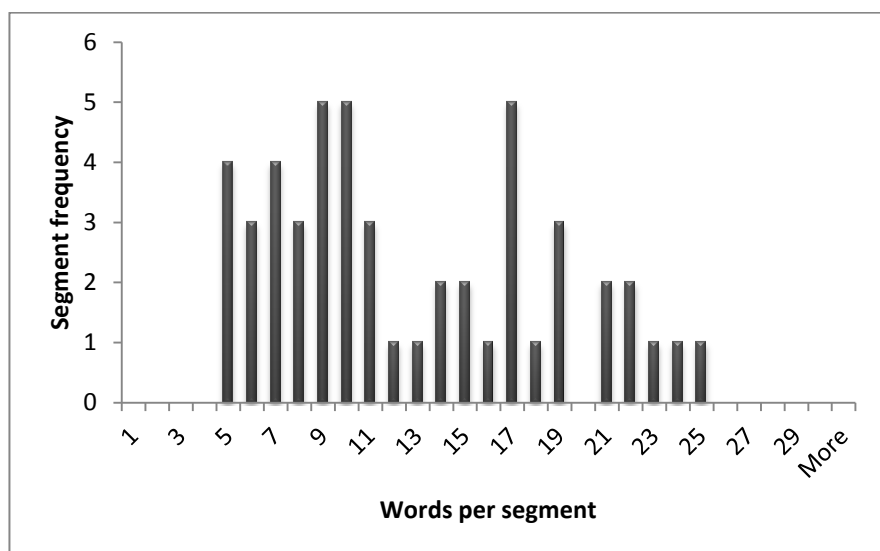


Figure 4: No match histogram

We wanted to make sure that there were no formatting changes as part of the fuzzy segments. Since the tool does not reflect format changes, the segments would appear as 100 percent matches to the translator instead of fuzzy matches. We had to make sure that the fuzzy matches were real fuzzy matches for the translators. Figure 5 shows the workflow for creating the dataset.

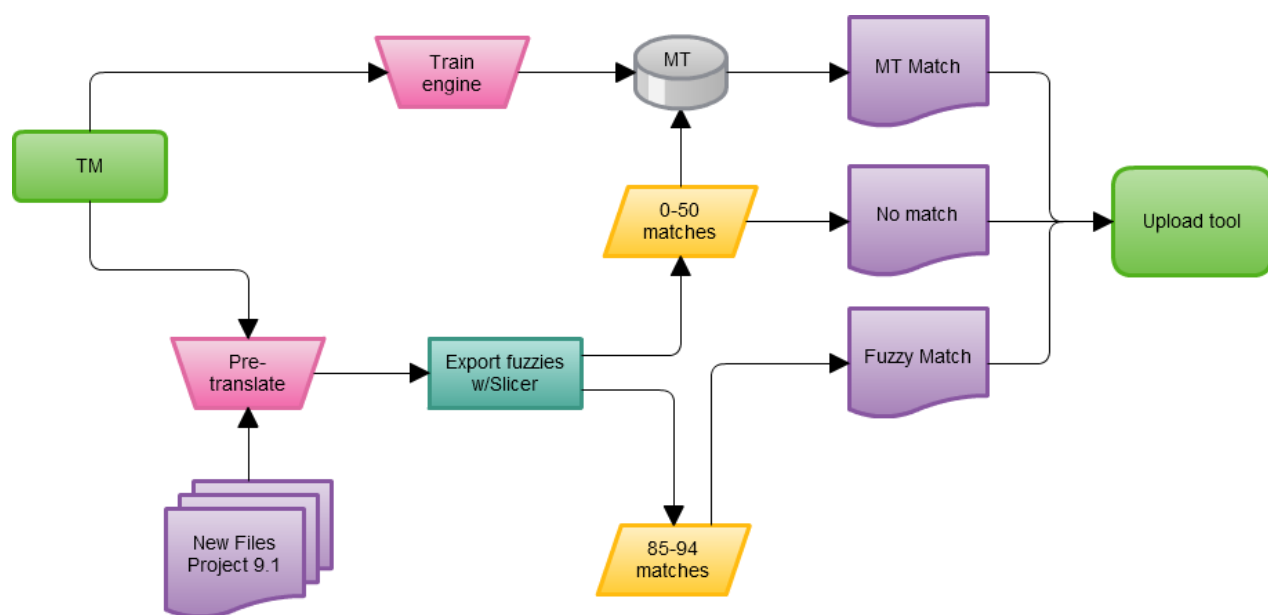


Figure 5: Dataset creation

Three initial strings (one of each category) were included for translators to practice on and become familiar with the tool, the glossary and the instructions. These were not included when measuring productivity or quality.

3.4.3.6. Post-editing tool

In order to measure the actual times and output produced by the post-editor we used a post-editing tool created by CrossLang. This tool is a web-based post-editing tool designed for post-editors; it enables customers to see the usefulness of machine-translated segments. The post-editors can connect online and translate or post-edit the proposed segments of text without knowing their origin (MT, TM or No match segments) and the tool measures the time taken in seconds for each segment. The tool is used to show customers how MT can be used to increase productivity in the translation phase.

This tool is an ideal instrument to test the defined hypotheses; since post-editors can post-edit and translate without necessarily knowing the origin of the text, and the tool can measure the time automatically as well as record the target texts. Further, it is not necessary, unlike with other methods, to have all post-editors monitored and measured in one specific location. They can work from home, anywhere in the world, in a familiar and relaxed environment, as with a standard translation assignment. Although this is not a standard CAT tool used by translators, it is not too different from other tools available in the industry (such as Google Translator Toolkit) in that it presents two

windows with source and target texts. Plitt and Masselot (2010) also use a specific post-editing environment very similar to this one “to measure time as precisely as possible” (ibid: 9). De Sutter and Depraetere (2012) and De Sutter (2012) have used this exact same tool for the experiments they have conducted. As one of the main aims of the study is to measure time, the use of other methodologies such as Think Aloud Protocols was discarded as it has been shown in several experiments that it slows down the translation process (Krings 2001, O’Brien 2006b). Finally, another factor is the financial aspect, as opposed to eye-tracking equipment, the tool is accessible and inexpensive.

The post-editing tool sends an email to the translators alerting them that a productivity task has been assigned to them. They are asked to click on a link and to make sure they keep the e-mail until they have completed all segments in this assignment. Once they have clicked on the link, they are presented with a screen containing the actual task as seen in Figure 6.

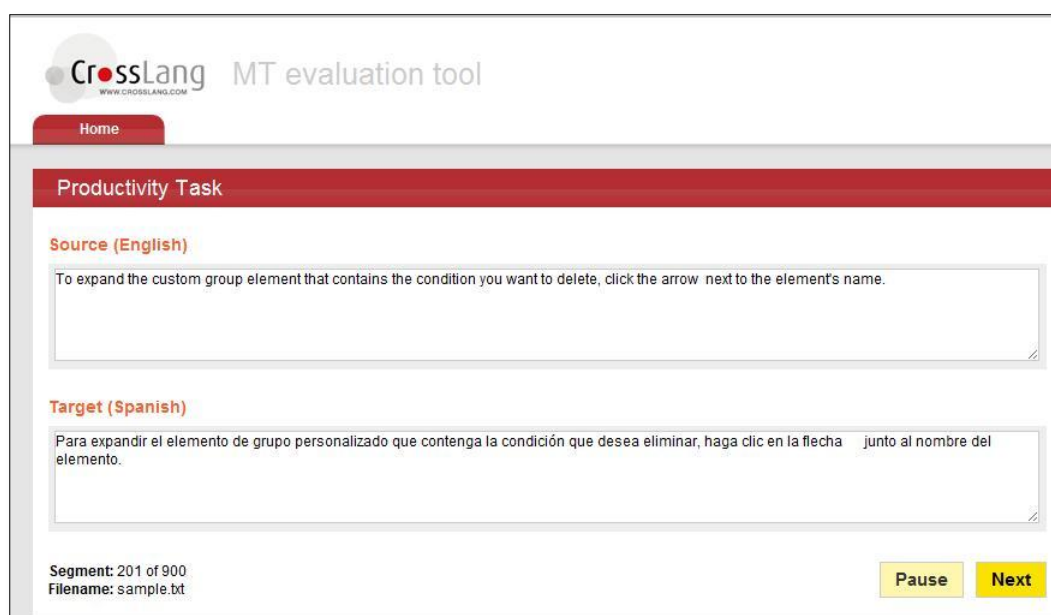


Figure 6: Screen-shot of post-editing user interface used

The Source window contains the source text in English, and the Target window contains either a blank text box or a proposed text in Spanish. The Spanish text is either an MT or TM segment. The post-editors are not aware of the actual origin of the translated text but they are asked to post-edit or translate the text. Once one string is done, the post-editor has to click on the Next button and proceed to the following

segment until they reach the end of the assignment. At this point, they are returned to the main window where they can log off.

Once the translators were finished, the team at CrossLang exported all the data from translators, containing translator name, segment ID, Source text, MT target text, Post-edited text, TM data, MT/PE difference (%), Segment length, Fuzzy match level, Post-edit time in ms, and Segment Origin.

3.4.3.7. Assignment instructions for translators

The translators received a set of instructions by e-mail explaining exactly the steps they needed to take to translate, edit and post-edit the segments. The instructions included how to interact with the tool, carry out the assignment, translate software options, follow specific guidelines on style, and how to use the Excel glossary provided. The translators were specifically asked not to stop unless strictly necessary, since the assignment was short, but they were given instructions on how to pause the assignment if they needed to. There was also a section on the quality expected. The request was: publishable quality, meaning full accuracy and no mistranslations with regards to the English text, compliance to Spanish language rules of grammar and spelling, compliance to the terminology following the glossary provided, and compliance to style according to the style guidelines provided (see Appendix C). The instructions also clarified that there should be no deletions or omissions in the text. Therefore, the translators had to make sure that they edited the text until the Spanish text was equivalent to the English text. However, they were advised to do this with as few edits as possible. They were asked not to introduce “preferential changes”, that is, they were asked to correct errors or make changes that they were certain about and that were fully justifiable according to the guidelines provided.

Because the tool did not allow the translators to review the segments once they had clicked the Next button, they were reminded in the guidelines that they had to review each segment fully before going to the next segment, as they would not be able to go back once they had submitted their proposals. They were also reminded that there was no spellchecker, so attention should be paid to typographical errors. Appendix C shows the exact instructions sent to the translators.

3.4.3.8. Assignment instructions for reviewers:

The instructions for the reviewers were more complex than for the translators, as the task involved taking an approach different from that performed during a normal review.

They were informed that they would have 24 individual translations in Word (24 translated versions of the same source text), 24 LISA forms in Excel to complete, and a timesheet in Excel in which to enter the time invested in the review task. They were also informed that they would need to review each translation in Word using the Track Changes option and then reflect the number of errors per category in the LISA form, one per translator.

Each Word document contained seven columns: Segment ID, Source Segment, Target Segment, Type of Match, Post-edited target, PE difference in %, and Type of error. The Target Segment could be blank if translators were not given a proposal and they had to fully translate the English segment. The reviewers were instructed that if the same error occurred throughout the translation of one participant, it should be counted only once (see section 5.2.4). The reviewers were informed about the Type of Match (Fuzzy, MT or No match) because they were requested to mark any overcorrection (see Overcorrection in Appendix A) found, although they were not to consider them as errors.

To track the time, the reviewers were given an Excel timesheet with two separate columns, one to include the time invested in correcting the Word document and another one to include the time invested in filling in the LISA form. They were told that they could go back and add more time if they realized, after completing the first review, that they wanted to change a correction or correct the text even further. They were also informed that it would be logical to have different times invested per translator because they were correcting the same text repeatedly. Since we wanted real-time data, it was not necessary to change the times to make them consistent for all texts. In the case of the reviewers, we did not have the possibility to time them automatically because of the reviewing method applied and the lack of an appropriate on-line tool. These data would be informative as the focus of our study was to measure the translators' productivity when using MT and fuzzy matches and not the reviewers' throughput.

There was a separate section to explain the review process in itself. The reviewers had to read the source text to understand the changes to make between proposed source segment and the proposed translation. Then, they had to read the post-edited target text to make sure it reflected the changes that would match the source text. They were also advised not to insert any preferential change (to only correct errors) to count the same error across a translation only once, to make minimal numbers of changes, to make sure the translators followed the glossary provided, and to be consistent across the 24

translations. Finally, they were given some clarifications on error typology according to the LISA QA model. Reviewers were asked to mark any overcorrection spotted in the target post-edited text with respect to the proposed segment, but not to count them as errors. They were also asked to be flexible as their ability to spot errors was not being judged, but the quality of the final translation was. There were separate instructions on how to use the LISA form. Finally, they were given the same quality expectations, style and terminology instructions as the translators. Appendix C shows the exact instructions sent to the reviewers.

3.4.3.9. The LISA QA Model

We used the LISA form to count and classify the number of errors (Appendix E) as in the pilot project and in De Almeida and O'Brien (2010). The standard takes into account the number of words in a given sample, and allows a percentage of errors in this sample (for example, one percent). The text might Pass or Fail the quality metric if the number of errors exceeds the percentage of errors allowed for that particular number of words. The focus was on the number and classification of errors, as the scope of this study was not to establish if a particular translator offered a good or poor performance (Pass or Fail) but whether the number and type of errors were affected by the use of a translation tool and therefore if the errors had an impact on the overall productivity of the translation. In other words, we needed to establish whether the time saved using MT or TM did not mean additional time in order to fix errors at a later stage in the localization process.

We introduced three categories under the type of error: No match, MT match or Fuzzy match segments. This was so that reviewers could insert each error in the category where the error was found so we could test our second hypothesis.

3.4.3.10. Translation Edit Rate (TER)

The Translation Edit Rate (Snover et al. 2006) (see also Translation Edit Rate (TER) in Appendix A) is a number that reflects the number of edits needed to change a hypothesis so that it exactly matches one of the references provided, normalized by the average length of the references. TER equals the number of edits divided by the average number of reference words. For us, the hypothesis was the post-edited segment (MT or Fuzzy match post-edited) and the reference was the proposed segment (proposed MT or Fuzzy match).

Possible edits include the insertion, deletion, and substitution of single words, as well as shifts of word sequences. A shift moves a contiguous sequence of words within the proposed translation to another location within the proposed translation. All edits, including shifts of any number of words, by any distance, have equal cost. In addition, punctuation tokens are treated as normal words and wrong capitalization is counted as an edit. This example illustrates the way TER calculates the score:

Sentence ID: 12

Best Reference: El texto automático información se mostrará dentro del campo de texto.

Original Hypothesis: La información de texto automático se mostrará dentro del campo de texto.

HYP: La @ de texto automático [información] se mostrará dentro del campo de texto.

EVAL: I S

SHFT: 1 1 1

TER Score: 27.27 (3.0/ 11.0)

In the text above, we have a reference sentence (the MT match provided to the translators): “*El texto automático información se mostrará dentro del campo de texto*”, and a hypothesis (the post-edited text): “*La información de texto automático se mostrará dentro del campo de texto*”. We see here that there is one insertion (“*La*”), one shift (“*información*” has been moved three words to the right) and one substitution (the preposition “*de*” instead of “*El*”). These are three changes in a total of 11 words (reference). Since the TER score is the number of edits divided by the average number of reference words, the result in this case is 0.2727 or 27.27 percent. This does not mean that the target text (hypothesis) is correctly post-edited; it only gives a score for the number of edits performed.

We decided to use TER because we wanted to investigate if the number of edits made when translating MT segments was similar to that when editing TM segments (fuzzy matches of the range selected) at a similar speed, and because research has shown that this metric correlates well with actual post-editing effort (Snover et al. 2006, He et al. 2010a, He et al. 2010b, O’Brien 2011, Offersgaard et al. 2008, among others). If we measured the time and also TER, we could see if the segments that took longer to edit were the ones that had more changes, and also if changes in MT and Fuzzy matches were similar in number and nature.

3.4.3.11. Glossary

The translators and reviewers were sent a glossary in Excel format with two columns, one containing the English source terms and the other the Spanish translations. The glossary had 703 entries. It was originally created by HiSoft and MicroStrategy to work on real projects and it contained core terms and software options. The glossary was then reviewed to include all references found in the actual project and some entries were deleted to avoid confusing instructions or ambiguity in the translation of certain terms. For example, the original glossary contained a complete list of products that had to be maintained in English in the translation but that did not appear in our project.

The translators were advised in the instructions that all the terminology needed was to be found in the glossary and they were advised to have it open during the assignment. They were also informed that it contained entries for both general terminology and software options and that these translations had priority over the proposed segments. This meant that if the proposed segments used different terminology or software options, the translators should follow the glossary. They were instructed to create their own translations based on the existing glossary if they did not find a particular term.

3.4.4. Data for qualitative analysis

This section includes the data used for the qualitative analysis although the first part of the questionnaire was used in the quantitative analysis to correlate translators' experience with productivity and quality.

3.4.4.1. Translators' questionnaire

The aim of the questionnaire was to establish the translators' experience in localization, tools, subject matter, post-editing, tasks performed, average daily throughput, and typing speed, in order to correlate this experience with their speed and number of errors. The results are presented in Chapter 6:. Moreover, we wanted to gather information on their opinions of revision procedures, post-editing, rates and the type of work they did, as well as on their work methodology. Finally, we wanted to know their opinions of the assignment, the tool used, the usefulness of the proposed segments, on the review process, and of how they worked with the terminology (glossary and text in the proposals). This data are presented in Chapter 7:.

SurveyMonkey was the tool used to publish the questionnaire. SurveyMonkey is a web-based survey solution that allows a survey to be created on-line and, sent to multiple participants, and allows the data to be collected in different formats. It also offers a summary of the data collected. A link was sent to the translators upon completion of the assignment, so this was a self-administered survey. The first page contained a presentation thanking the translators for participating in the questionnaire and telling them that the questions would take no more than 30 minutes. The first page also stated the aims of the study and named the bodies funding the research. We decided to start with questions about their experience in order to ease the participants into the study, avoiding awkward questions (O'Leary 2010), and because we would use the section on experience to cross-reference it with productivity and quality. This first part of the questionnaire was descriptive. We left the questions on their opinions to the second part, followed by the questions on the assignment. These parts were designed to gather qualitative data to potentially understand or explain better the quantitative results. The demographics were left for the final part, since they were not the focus of the research but we believed it could be important to have this information to relate experience with age and technology (since translators with more experience and who are older might not be as exposed to technology as is the younger generation of translators). The questionnaire consisted of 28 questions that address these aspects. The questions were presented so that they would be intuitive and they would follow a logical flow (Holyk 2008). There were multiple choices with only one answer, multiple choices with multiple answers, rating scale questions to rate relevance in opinions, a matrix of choices and open text boxes for their personal comments. Appendix D shows the questionnaire received by the translators.

3.4.4.2. Reviewers' questionnaire

The aim of the questionnaire was to define more specifically the reviewers' prior experience, even though the selecting criteria were quite specific. We wanted to have the exact number of years of experience in each field and also to gather their opinions about their review work, machine translation and the assignment. The questionnaire was very similar to that of the translators, but with more emphasis on the reviewing aspect. It was also divided into three sections that covered their experience, their opinions about work practices, and their opinions of the reviewing assignment. Appendix D shows the questionnaire received by the reviewers. The results are presented in Chapter 7:.

3.4.4.3. Debriefings with translators and reviewers

After the translators and reviewers completed the questionnaire, we asked them if they would accept to take part in a one-on-one discussion about the assignment that would be recorded. We used Skype, an internet application that allows internet calling, since our participants were located in different cities in the globe. Two applications were used to record the conversations: Pamela Call Recorder for Skype and Pretty May Call Recorder for Skype (we used two just in case one of them did not work at any given moment with any of the participants).

The objective of the one-on-one interview was to add qualitative data that would help explain the quantitative data collected. One-on-one interviews allow “the interviewee the freedom to express their thoughts” (O’Leary 2004: 164) and this might be curtailed in a group. The data would be gathered immediately after finishing the assignment so the ideas could be fresh in their minds. Despite the fact that the debriefings took place before we had results from the quantitative analysis and therefore we did not have specific questions about particular issues, a discussion immediately after the assignment could help clarify quantitative issues found at a later stage. If we waited to analyze the quantitative data, times or errors, the participants might not remember why they performed a certain action or what their general ideas were about the assignment. Therefore, we decided to use an informal semi-structured interview with a flexible structure. We wanted to have some questions to start up the conversation but these questions had to be flexible to allow participants to make personal comments. We were looking to elicit opinions, feelings and thoughts about the assignment, machine translation, rates and their review process.

The debriefings began with a short introduction on the study, afterwards participants were reminded that they were being recorded, how the data would be treated and that they were not obliged to respond to the questions if they did not feel they were appropriate. There were seven questions for the translators in Spanish. We formulated the questions in Spanish so that the participants would feel comfortable and relaxed. English might be their working language, but Spanish, as their native language, would allow a certain ease when expressing an opinion or feeling. The interviews were translated and transcribed into English (see Appendix I and Appendix J). The questions were:

- What did you think of the instructions for the task including the glossary?

- Did you know in advance that this was a project containing MT segments? Did you think about it at any particular moment during the assignment? How did you know?
- Did you notice any difference between the proposed segments? Do you have any examples?
- Was there any segment that you found more difficult to translate or edit? Why?
- Which questions in the questionnaire were most difficult to answer?
- How did you feel doing the task?
- Would you like to add any comment?

There were ten questions for the reviewers:

- What did you think of the instructions for the task including the glossary?
- How did you find the review methodology proposed?
- How do you normally review translations?
- Which errors were difficult to classify?
- Of the 24 translations, were there any that caught your attention for a particular reason?
- Did you find that the translations were similar with respect to the type of errors made? Did you find a lot of over corrections?
- Was there any segment that was difficult to translate? Why?
- Which questions in the questionnaire were difficult to answer?
- How did you feel when completing the task (while working in the assignment)?
- Would you like to add anything else?

The questions aimed at eliciting a conversation with the participants and listening to what they might have omitted during the assignment or we might have not asked in the questionnaire, and to see where they would place particular emphasis when giving an opinion or where they had found difficulties in the assignment or even in the translation process or the localization industry. These questions gave us a framework from which to derive recurring themes. We used NVivo 9.0 to code and analyze the debriefings. NVivo allows you to automatically code the questions asked making it easy to organize the questions with all responses from the 23 participants.

The following is the initial framework designed for the translators:

Integration between MT and TM

1. Assignment
 - 1.1 Instructions
 - 1.2 Glossary
 - 1.3 Questionnaire
 - 1.4 Tool
 - 1.5 Segments
2. Machine Translation
 - 2.1 Feelings: Positive, Negative, Mixed
 - 2.2 Knowledge of MT processes
3. Translation Processes

And for reviewers:

Review Process

1. Assignment
 - 1.1 Instructions
 - 1.2 Glossary
 - 1.3 Questionnaire
 - 1.4 Methodology
 - 1.5 Segments
2. Machine Translation
 - 2.1 Attitude: Positive, Negative, Mixed
3. Review Process

3.4.5. Validity and generalizability of the research

As seen in the literature review the number of participants in post-editing studies tends to be small (fewer than 10 participants). Our sample has 24 translators and three reviewers (also translators) and, could thus be considered a relatively large sample. Further, the sample text used contains 2,124 words in 149 segments, which generated 3,576 observations from 24 translators and 10,728 from three reviewers for the statistical analysis, giving us sufficient data to establish the statistical significance of our results and their generalizability.

The validity of the research is also enhanced by a series of features that are either taken directly from professional practice or that are as close to professional practice as the experiment design would allow.

A language service provider (LSP) selected the participants according to their availability and criteria established. The participants, both translators and reviewers, were all professional freelance translators (English into Spanish language combination) existing in the LSP's freelance database. This happened in exactly the same way as in any other localization project. They were also paid for the task following the standard channels: they received a purchase order, they invoice, and they were paid. They worked from their home or office in their own authentic environment and started the task after receiving relevant instructions by email. Although, the participants were aware that the assignment was part of a research project (and they signed a Research Participant Release Form), the methods for contacting them and assigning the project were the standard methods in the workplace.

The engine was trained with authentic translations memories from one specific customer. The output was evaluated automatically and also by a human translator in order to establish the existing quality, using the same evaluation tools as in a commercial setting. Therefore, the standard procedures were used to assess the quality of the output and this output was not modified to favor the aims of the study.

The source text was authentic and it was only manipulated to select the relevant number of words and fuzzy matches necessary for the experiment. We made sure that the segment length in the sample reflected the actual content of the whole project in order to mimic a real life situation (see section 3.4.3.5). Using the structure of the original corpus, the representative segments were applied, so that our sample contained the same segment length as the original corpus (see Figure 4). This means that the source text had the same characteristics as a standard localization project for this particular customer. As to the volume of work, the standard number of words used to plan projects is 2,500 “new” words a day in the localization industry. We were using a little above 2,000 words (in different types of segments to process since in our case not all of the words are new), which represents between 35 and 40 percent of a working day and this, compared with similar studies of the same nature, is a reliable volume to test our hypotheses.

The on-line post-editing tool was specifically designed to test translators' productivity using, if need be, different text proposals. Therefore, the tool measures the

time taken to complete a given number of words accurately without having to rely on the participants' intervention, and moreover it disguises the origin of the segments (whether or not they come from MT) thus avoiding any bias from translators with respect to the origin of the segments. This gives validity to our measurements. The tool requires no installation, but only a simple connection on-line, and it is simple to use. Although this is not a standard CAT tool, it shares common characteristics (e.g. the tool provides one window with a text box containing the source text and another text box, just below it, that is either blank or has a translation proposal).

The reviewers used the LISA QA Model, which is a model frequently used in the localization industry to assess translators' quality, and this meant that reviewers had to be familiar with the criteria and so that the lack of knowledge of the model would not act as a confounding variable. Having three independent reviewers diminished the possibility of bias or personal effect in the final data.

The questionnaire addressed the participants' experience in more detail, through a series of questions that gave us a comprehensive overview of each translator's experience. Further, the second and third parts of the questionnaires, complemented by post-task debriefings, provided more information on the difficulties and feelings of the participants during the assignment, and this data helped us explain the results obtained during the quantitative analysis, and establishes if there are further threats to the validity of the assignment. The questionnaire was administered using an on-line survey tool that facilitates access to translators and data gathering for the researcher. The voluntary debriefings were conducted from the comfort of the participants' homes and in their native language, and it served to triangulate the data with the results obtained during the survey.

Finally, the statistical tests selected and the professionals performing this analysis guaranteed the reliability and validity of the statistical calculation and results.

3.4.6. Threats to validity

We have seen in the section above the measures taken to ensure the validity and generalizability of our experiment. However, we need to make certain considerations about possible threats to validity.

3.4.6.1. Languages used

The language combination used is English to Spanish. This language combination has the advantage of being very widely used in the localization industry and therefore the conclusions may be useful to a wider community. For this same reason and because of the amount of parallel corpora in this language combination, machine translation engines tend to work better in this combination than in others. It is thus not advisable to extrapolate directly any result to other language combinations. We selected this language combination mainly because it is popular in the localization industry (if a product is to be translated, Spanish is normally one of the target languages chosen) and it is therefore of interest to the community, it has been widely used in machine translation, and it is the mother tongue of the researcher. This makes it easier to train an engine with sufficient bilingual data, select the texts, find initial mistakes in the corpus, test all samples and look at the final post-editors and reviewers' target texts. Moreover, we have a limited budget that does not allow paying for additional languages.

3.4.6.2. Language variant

This project was setup with the Vendor Management team at HiSoft as a standard localization project, even though participants were informed that this was a research project. Therefore, translators from the English-to-Spanish language combination were contacted regardless of their country of origin, as was normally done in this company when the target text required was not constrained to one particular country. The glossary and the technical nature of the text contributed to the avoidance of misunderstandings in language usage: the different varieties of Spanish were not an issue here.

3.4.6.3. Selection of post-editing tool

The post-editing tool chosen was appropriate for the experiment. It was easy to use, free and required practically no training. In addition, there was no need to carry out the experiment in a laboratory with special equipment or on-site. Despite these advantages, it had some shortcomings: the translators could not go back and correct a segment if they realized they had made a mistake, which often happens in a real-life scenario. We compensated for this by creating very specific instructions explaining the functioning of the tool and emphasizing the need for a complete review after processing each segment and before hitting the Next button. However, we are aware that even in this case translators often go back and revise their work, realizing only afterwards that they made a mistake at the beginning of the assignment. Finally, post-editors do not use this as a

standard tool and although it does resemble some of the current on-line tools they might use, we realize that others, such as SDL Trados, have more options, with the corresponding advantages and disadvantages. Therefore, the times and errors were not measured in their natural environment. Testing directly using a CAT tool might have given us different times, since other actions needed to be performed, such as opening and closing the segments. The alternative of asking translators to measure their own times while working with a CAT tool was discarded as each translator might measure time differently, even despite the fact that specific instructions are given. Although “time stamps” are recorded in SDL Trados 2007, “this function is somewhat problematic” (see Tatsumi 2010: 67), as the application overwrites the saved times after the second time, and it does not record the time during which a segment is open. Further, the current tool did not inform the translators about the origin of the segment and thus it did not create a bias towards any of the proposed segments, and this was of interest for our research.

3.4.6.4. Revision by a third party

We used three reviewers to review 24 translators. They had to correct the translations using the Track Changes function in Word and then document the changes in the LISA form in Excel. The process was slightly more cumbersome than if they had to simply implement changes in the files and give an overall score. Also, they reviewed the same text twenty-four times, making the task similar to a translation test evaluation rather than reviewing a live project. Finally, the reviewers measured their time at home using an Excel form, which we realized might not reflect the exact time invested per file.

3.4.6.5. Statistical engine

A statistical engine was used for the project. We are aware that the results could be different with different engines, but budget limitations forced us to focus on one particular engine.

3.4.6.6. Sequence of segments

Since we wanted to test three match categories (No match, Fuzzy match and MT match) and a certain type of fuzzy match (85-95 percent), we could not choose consecutive segments because they might not have had the desired level of match. We selected all segments from this level of fuzzy match from a large base of segments and then uploaded them to the post-editing tool. We understand that this is not the ideal

composition of a text. Nevertheless, post-editors are frequently asked to focus on certain segment types while skipping others, losing the macro-context and having to concentrate on the micro-level of isolated segments. So texts are not translated in a linear sequence. Thus this type of segmentation is quite frequent in a localization project.

3.4.6.7. Selection of translation memory system

We used SDL Trados 7.1 to extract the translation memory segments. We chose this tool for two reasons: firstly it is the most commonly used tool in the industry (Lagoudaki 2006), and secondly the corpus we are using had been created using this tool. We understand that each CAT tool defines levels of fuzzy match differently because they use different algorithms, but testing several CAT tools was beyond the scope of this project.

3.4.7. Testing the methodology

One participant carried out a first initial test of the project between May 18th and 24th 2010. We presented the participant with a set of instructions and a core glossary by email. The participant was instructed to carefully read the instructions in PDF format and was then sent a link to access the post-editing tool. No further instructions or training were given for the tool. Upon finishing the test, the participant was sent a link to the on-line SurveyMonkey questionnaire to complete. The test proved to work without any incident and the participant understood the tasks and carried them out as initially planned. This first translator was also debriefed in order to see if the instructions were clear and if changes needed to be applied. It was interesting during this phase to see that, although she had worked faster with machine translation, her impression was that she had been faster when typing her own translation because, as she mentioned, she was a very fast typist. The questionnaire was completed successfully and with the input received we implemented changes to clarify certain questions. The post-edited segments were then sent to three reviewers. They followed the instructions, sent the corrections back and completed the questionnaire. Once they had finished they were interviewed. They found the instructions easy to follow but there were certain unclear aspects. For example, they did not know if they had to count the repeated errors more than once, they found one inconsistency in the glossary, and some options that could not be selected in the questionnaire.

After this initial test run, changes were implemented to correct the issues that arose. Since we saw that the reviewers had inserted some preferential changes (see Preferential changes in Appendix A), we modified the instructions so they would more clearly specify to reviewers that they had to correct only errors and not insert preferential changes.

3.5. Project development

3.5.1. Translation phase

Once the testing phase was finished and changes were implemented in the instructions, the questionnaire and the glossary, we contacted the rest of the translators. They were contacted in an initial email on May 31st 2011 with specific instructions. They were informed that the project would start on June 2nd 2011 at 10 am (Spanish time) and end on June 3rd at 6 pm (Spanish time) and that they would receive the instructions and the task. They were advised that it was important to print and read the instructions carefully before starting. There was also a brief description of the task so they would be prepared and could organize their day's work. They were also asked to contact us if they had doubts and to confirm receipt of the message.

We had to replace two translators who sent an email stating that they were familiar with the researcher's publications on post-editing. We thought that this familiarity could distort the results and that it would be better to contact translators who might not be influenced by previous publications on a similar project.

We sent out the instructions and glossary finally on June 1st at 2 pm. In the email the translators were informed that they would receive another link to access the assignment. They were reminded again to print the instructions and to click the Pause button if they had doubts and wanted to ask questions. They were warned, as well, about the importance of doing the assignment without stopping within the same day unless strictly necessary. They were also asked to contact us once they were finished so they could fill in the questionnaire and answer some questions through Skype. No further instructions or further training was given on the tool or on the nature and type of segments they were to complete.

There were several queries and issues during the start-up phase. We think these are relevant to see if the issues could be related to the final productivity and quality

obtained. In other words, can we tell the outcome depending on the types of issues experienced by the translator or the types of queries made beforehand?

Translator 3 asked which second person singular to use when addressing the reader: the formal “usted” or informal “tú”. This translator was told to use the formal “usted” as in regular IT assignments.

Translator 17 deleted the initial link thinking that it might contain a virus but then recovered it.

Translator 10 opened the link and started the assignment without reading the instructions or following the glossary. Once she realized this, she contacted us. Since there were three testing segments, this translator was informed to continue but now following instructions and glossaries.

Translators 15, 16, 17, and 20 asked about scheduling and the possibility of stopping the assignment. They were informed that it was better to do it in one go due to the size of the assignment.

Translator 20 asked for confirmation on the terminology process: if indeed they had to check every term of the new and proposed segments in the glossary because this would obviously add time. This translator was informed that they did have to follow the glossary and that we were aware this would mean more time at the end of the process.

Translator 16 asked about the quality expected and if he should stop in case he had to check cross-references. If they were required to achieve publishable quality but at the same time they could not use spellcheckers, print the text, and they did not have context, and they were not able to go back, accomplishing this quality was more difficult. This translator was informed that within the characteristics of the project, this was the quality to be expected. He was also informed that there were no cross-references.

Translators 3 and 21 asked about the instruction on pausing and losing changes in the post-edited segment. We clarified the process.

Translator 22 sent a message to CrossLang saying that she did not know what to do with the link sent to her. The Vendor Management team contacted her to clarify that she had accepted to take part in the job and she received instructions.

Translator 17 asked about how to report errors in the source text. She was informed that she could report errors in a document.

In view of these queries, we sent another general email informing the translators that:

- They could not respond to the automatic link since nobody would respond. They could address any queries to us.
- The glossary was to be consulted during the translation without pausing the assignment. They were advised to have it open and to check it at the same time.
- The ideal procedure was to complete the assignment in one go but if they had to get up or if they had a call, they could click Pause. Also, they were informed that if they turned off the computer they would not lose their work.
- The questionnaire would be sent after completing the assignment.
- If they did not have Skype or they did not want to use it, this was not a problem.

3.5.2. Schedule

Table 5 shows the start and end times (in the format MM/DD/YEAR), and locations of all 27 participants. The column Instructions shows the date when translators confirmed they had received the instructions, not when they were sent. The column Task reflects when the assignment was finished, not when it started. The columns Questionnaire and Debriefing show when the tasks were started and completed.

Translator	Instructions	Task	Questionnaire	Debriefing	City	Country
TR01	06/01/2011	06/02/2011	06/02/2011	06/06/2011	Buenos Aires	Argentina
TR02	06/02/2011	06/02/2011	06/02/2011	N/A	Barcelona	Spain
TR03	06/01/2011	06/02/2011	06/03/2011	N/A	Córdoba	Argentina
TR04	06/01/2011	06/02/2011	06/03/2011	06/03/2011	Santa Fe	Argentina
TR05	06/01/2011	06/02/2011	06/02/2011	06/02/2011	Granada	Spain
TR06	06/01/2011	06/02/2011	06/02/2011	06/08/2011	Barcelona	Spain
TR07	06/01/2011	06/03/2011	06/03/2011	06/06/2011	Buenos Aires	Argentina
TR08	06/01/2011	06/03/2011	06/03/2011	06/08/2011	Córdoba	Argentina
TR09	06/02/2011	06/02/2011	06/02/2011	06/03/2011	Altafulla	Spain
TR10	06/02/2011	06/02/2011	06/02/2011	N/A	Rosario	Argentina
TR11	06/01/2011	06/02/2011	06/03/2011	06/03/2011	Buenos Aires	Argentina
TR12	06/01/2011	06/03/2011	06/03/2011	06/07/2011	Madrid	Spain
TR13	06/01/2011	06/03/2011	06/03/2011	N/A	La Plata	Argentina
TR14	06/01/2011	06/03/2011	06/03/2011	06/03/2011	Rosario	Argentina
TR15	06/01/2011	06/02/2011	06/03/2011	06/03/2011	Buenos Aires	Argentina
TR16	06/02/2011	06/02/2011	06/02/2011	N/A	Cádiz	Spain
TR17	06/02/2011	06/02/2011	06/02/2011	06/02/2011	Buenos Aires	Argentina
TR18	06/02/2011	06/03/2011	06/03/2011	06/03/2011	Buenos Aires	Argentina
TR19	06/01/2011	06/02/2011	06/02/2011	06/06/2011	Sydney	Australia
TR20	06/02/2011	06/02/2011	06/02/2011	06/02/2011	Barcelona	Spain
TR21	06/01/2011	06/02/2011	06/02/2011	06/10/2011	Madrid	Spain
TR22	06/01/2011	06/02/2011	06/03/2011	06/06/2011	Santiago	Chile

Translator	Instructions	Task	Questionnaire	Debriefing	City	Country
TR23	06/01/2011	06/02/2011	06/02/2011	06/02/2011	Barcelona	Spain
TR24	06/01/2011	06/02/2011	06/02/2011	06/08/2011	Madrid	Spain
REV01	06/07/2011	06/14/2011	06/14/2011	06/14/2011	Córdoba	Argentina
REV02	06/07/2011	06/14/2011	06/14/2011	06/14/2011	Cáceres	Spain
REV03	06/07/2011	06/14/2011	06/14/2011	06/14/2011	Buenos Aires	Argentina

Table 5: Assignment dates and locations

There were issues with two participants that had problems selecting the Next button. Translator 9 stated that she had problems with Google web browser and that she hit Next twice without editing two segments; TR04 said that she had a problem with the power supply and that she had lost three segments.

Translator 11 included some comments in his email upon completion of the tasks. He had made a mistake in two segments and realized this once he had hit the Next button, so he sent an email to explain. We sent this information to the three reviewers.

Translator 22 did not confirm receipt of the task, so we had to contact her in order to see if she was finally able to work on the project. She then confirmed.

Translator 13 did not notify us when she had finalized the task, in order to send the questionnaire. Once we asked her if she had finished, she informed us that she could take the questionnaire.

All translators except Translators 2, 3, 10, 13 and 16 took part in the debriefings, as did all three reviewers. There were different reasons given for not participating: Translator 2 did not want to install Skype, Translator 13 did not have good internet connection because she lived in the countryside, Translator 3's microphone was not working properly, Translator 16 did not want to do it, and finally Translator 10 was not contactable after completing the questionnaire (she was on holidays for two weeks immediately afterwards). Since we wanted the interview to be a place where the translators could express themselves freely and give detailed information, we did not want to make it a compulsory part of the research.

3.5.3. Reviewing phase

Once the testing phase was finished and changes were implemented in the instructions, the questionnaire and the glossary, the three reviewers were also contacted to let them know that they would receive 48,000 words on June 7th and that they would need to be finished by June 14th at 6 pm (Spanish time). They had previously been told that these words were in fact the same 2,000-word text repeated 24 times.

Once the translators finished their assignment, the reviewers were sent an email with the 24 numbered translations, 24 numbered review forms, a time tracker, the instructions sent to translators, the instructions for reviewers, and the comments from Translator 11. They were told to contact us once they had finished to complete a questionnaire and a short interview on Skype and to go over the instructions with care, since this was not a standard review. They were also asked to contact us if they had problems with the error classification.

The reviewers quickly confirmed the assignment and that they had understood the instructions. Reviewer 3 contacted us with questions related to software options and capitalization. The response was as follows: her criteria could be used to establish the correct translation, but we reminded her that the translators were instructed to follow the glossary as a first priority and that in this glossary the capitalization followed the standard Spanish norm. There were no more questions and the reviewers confirmed completion of the assignment on the due date. They were sent the questionnaire, which they completed without any issues. They were also interviewed on Skype upon completion of the whole assignment.

Once the process of uploading data commenced, it was noted that Reviewer 1 had not sent the Word documents with Track Changes as per instructions and Reviewer 2 had not specified the type of error in the Word documents. Reviewer 1 had placed the wrong files in the zip file when sending the order; Reviewer 2 had not inserted the type of error when doing the Track Changes and she had to do it again and add the additional times in the time tracker. In the case of Reviewer 3, some errors in Word were not rendered correctly in Excel, so she was asked to review the texts and make sure both documents matched. Similar issues were then found with Reviewer 1 and Reviewer 2, where some errors were marked in the Word files and not quantified in the Excel files. All reviewers had another chance to compare the files and make sure both forms matched. Information on error classification was exchanged as we were uploading the data into one Excel file. This information was mainly related to errors marked and not counted, and, on occasions, the reasons behind these decisions. This “query” time was not added to the time tracker. The process of reviewing was cumbersome because they had to mark the changes and type of errors in Word and then transfer them on to the Excel spreadsheet. Although the reviewers did understand the task correctly, judging by the initial emails, it was more difficult for them to produce consistent results. However, the reviewers were very proactive and willing to clarify doubts and correct their own

mistakes when rendering the results on the LISA forms. The description on how errors were compiled and calculated will be presented in the section on Quality.

In this chapter we have described the methodology used for the project explaining its validity and generalizability as well as possible limitations, we have also given a short description of the project development. In the following chapters we will present the results obtained.

PART II: Quantitative results

In this part we will present the quantitative results obtained from the project, including a brief description of the statistical analysis. Chapter 4 contains the results on productivity, including TER results and speed groups; Chapter 5 gives the results on quality, including the review process and the final errors, as well as the relationship with processing speed; Chapter 6 deals with the translators' experience, including the relationship with processing speed and errors.

II.1 Statistical analysis

During the project, we gathered data from the post-editing tool translators used (time), from the on-line questionnaire (experience) and from the reviewers (errors). We created several databases and we applied different statistical methods. Here, we will explain the content of the databases and secondly, the statistical methods applied.

From the 24 translators we obtained one database with the exported txt files. We compiled this data into an Excel file containing:

- The translators' identification numbers
- Segment identification: individual number per segment provided by the tool
- MT target: column containing both proposed MT matches and proposed Fuzzy matches
- Post-edited segment that showed the final translator's version for each translator
- Match category that indicated if the segment was MT match, Fuzzy match, No match
- Segment length: number of words per segment
- Fuzzy match score: score obtained from SDL Trados, and ranging from 85 to 94
- Post-edited time in milliseconds: time invested by translators in doing the task
- MT/PE difference (%): Olivier and Hand score (automatically provided by the tool) (Olivier and Hand 1996)

- TER: the TER for both MT and Fuzzy match segments (calculated by us and explained in section 3.4.3.10)
- Segment origin: reference to the ttx file that the text came from.

For the statistical analysis, one database was created with 3,576 registers (149 segments translated by 24 translators), and also all answers from the questionnaire with 24 registers (responses to the on-line questionnaire from the 24 translators).

From the reviewers we had 24 LISA forms, 24 edited Word documents (with tracked changes) and one timesheet noting the time employed to correct each text and to complete each form. With this information in hand, all errors were transferred to three Excel documents:

- One contains all the errors as they appeared in the LISA forms: Reviewer ID, Translator ID, Number of errors, Number of error points, Categories (No match, Fuzzy match and MT match), Error Type (Accuracy, Style, Terminology, Language, Mistranslation), Severity (Minor, Major, Critical). A second tab includes the overcorrections marked by reviewers (see Appendix A for a definition of Overcorrection).
- The other Excel document contains the same categories as the previous one as well as the Segment ID, since all errors per individual segment were transferred from the Word documents, and not the global amount of errors per text.
- The final Excel file contains the time invested by the reviewers and it includes: Translator ID, Time invested in the Word document, Time invested in completing the form, Total time and Reviewer ID.

For the statistical analysis, three databases were created:

- One database containing the text revision with 216 registers (24 translators * three reviewers * three Match categories) and 12 variables (Translator, Reviewer, Match category, Accuracy, Language, Mistranslation, Terminology, Style, Country, Format, Consistency, Total errors);
- One review database with all segments translated containing 10,728 registers (149 segments * 24 translators * three reviewers) and 16 variables (Segment, Translator, Reviewer, Match, Segment length,

Accuracy, Language, Mistranslation, Terminology, Style, Country, Format, Consistency, Number of errors, Total errors, Overcorrection number);

- One containing 72 registers (24 translators * three reviewers) and eight variables (Translator, Reviewer, Reviewer, Word document time, Form time, Total time in Words per minute, Words per minute in Word document, Words per minute in Form) containing the times spent during the revision by the three reviewers for the 24 translators.

With this data we carried out the statistical analysis together with the Servei d'Estadística Aplicada (SEA) of the Universitat Autònoma de Barcelona. SEA is a scientific and technical service aimed at supporting researchers. They are involved in numerous projects in all fields of science. The data reading, manipulation and validation were performed using the software SAS v9.2.

For the quantitative analyses, descriptive statistics with tables and graphs were used to summarize the relevant information about the data set, showing means, medians, standard deviations, minimum, maximum and missing values, as well as minimum, quartile 1, median, quartile 3, maximum, range and range quartiles. Different inferential statistical methods were used to draw conclusions from the observed data depending on the hypothesis we wanted to test. We describe the methods used within each area of this study.

II.1.1 Productivity

To compare the translation speed using different methods (MT, TM or No match), the continuous variables (see Appendix A) *Post-editing time* (total time to process the assignment) and *Words per minute* (number of words processed per minute) have been explored. To analyze these variables we defined a linear regression model with repeated measures (translator) to compare the response variable (*Post-editing time* and *Words per minute*) according to the three categories. In these cases, we applied logarithmic transformations to the response variable in order to obtain a normal distribution for variable.

The linear regression model was used with repeated measures (Littell, Stroup and Freund 2005) because the database contained 149 segments translated by each of the 24

translators. Thus we could identify all the segments that came from the same translator. In other words, the effect *translator* is introduced in the model.

To analyze if the initial speed group (taken from the No match segments) is a speed predictor for each translator when processing Fuzzy and MT match segments, we established a linear regression model with repeated measures taking *Standardized Words per minute with respect to No match* as the response variable. This variable was defined by standardizing the translation speed (Word per minute) for each translator according to their mean speed and the standard deviation in all their No match segments. In this model, we included the explanatory (independent) variables *Match category* (Fuzzy and MT) (see Appendix A) and a re-categorization (redefining the categories) of Words per minute.

To explore the edits made by translators in the different proposed translations, we looked at the TER metric (as explained in section 3.4.3.10). Firstly, we investigated the correlation of the selected index with the Levenshtein, and Olivier and Hand indexes, using dispersion diagrams and a Pearson correlation coefficient (which measures the degree to which two variables are linearly related) for each pair of indexes (see Figure 13 on page 107 for an example). Secondly, we created a TER indicator, a variable that showed if a segment was edited or not. To check if there were differences in this indicator among the different categories, Fuzzy and MT match, we created a binary logistic regression model with repeated measures (translator). Thirdly, we checked if the TER score showed statistically significant differences for MT and Fuzzy match. With this objective in mind, we created a linear regression model of repeated measures (translator) with the response variable *TER* and with the explanatory variable *Words per minute* in four categories (<10 words per minute, between 10 and 20 words per minute, between 20 and 30 words per minute, and more than 30 words per minute).

II.1.2 Quality

To analyze the differences in the revision times of three reviewers we used a linear regression model with repeated measures (translator) with the response variable *logarithm of Review time*. We used this model with repeated measures because the database had three revisions of the same translated text, and we wanted to look at the coincidences among the three reviewers for the same translators. To analyze the differences in errors of the three reviewers we categorized the error variable in three

categories according to the first and third quartiles, and we applied the Kappa statistical measure in order to see the level of similarity or agreement.

We defined an error indicator to compare the presence or absence of errors in each segment. The similarity among reviewers was analyzed by defining an indicative variable of disagreement between them. We applied a logistic regression model with repeated measures (translator and reviewer) taking *Error indicator* as the response variable and *Match category* as the explanatory variable. We used this model with repeated measures because the database has three revisions of the same translated text, and we wanted to introduce the effect of having these translators and reviewers.

We used the error counts from the global dataset (errors from all LISA forms) in order not to have the same repeated error included more than once. To model the error count, considering that it is a discrete variable, we established a Poisson regression model with repeated measures (translator and reviewer) with an offset of the text length. The objective for including the offset was to standardize the error count with respect to the total number of words of each segment. It is important to note that Severity, and thus error points, has not been considered in this analysis because most errors were classified as Minor errors, and only missing translations (as in Translators 4 and 9, see section 3.5) were considered Major (equivalent to 5 error points, see LISA and the LISA QA Model in Appendix A). Since the difference between errors and error points was so low and the severity applied was due to technical issues on the part of translators, we decided to use the number of errors count and not the error points.

The total error database was also used to model the *logarithm of Total time in words per minute* as the response variable and *Total errors* and *Reviewer* as explanatory variables. We used this model with repeated measures because the database has 24 translations of the same text.

We established a Poisson regression model with repeated measures (translator and reviewer) taking *Total errors* as the response variable and *Match category* and *Speed group* as the explanatory variable. All statistical decisions have been established with a 0.05 significance level.

II.1.3 Experience

In order to characterize similarity among translators on the basis of the answers from the questionnaire regarding their experience and to then see the differences between the

clusters so defined, a multiple correspondence analysis (Greenacre 2008) was setup. The statistical decisions were made taking a 0.1 significance level.

In order to see if there are differences between the groups defined in the multivariate analysis with respect to their translation speed, a linear regression model with repeated measures (translator) was applied with the response variable *logarithm of Words per minute* and the *Match category* and *Cluster* and the interaction between them as explanatory variables.

In order to see if there are differences among the clusters defined by the multivariate analysis with respect to translation errors, we applied a Poisson regression model with repeated measures and with the offset *text length* and the response variable *Total errors*, and with *Match category*, *Cluster* and the interaction between them as explanatory variables.

Chapter 4: Productivity results

In this chapter, we will look at data in relation to the first hypothesis, which claims that the time invested in post-editing machine-translated text will correspond to the time invested in editing fuzzy-matched text corresponding to the 85-94 percent range. We will also test the sub-hypothesis that translators with higher processing speeds will show less productivity gain when post-editing MT or Fuzzy match segments than translators with lower processing speeds when translating No match segments. Finally, edits made by translators in the different segments will be analyzed using the TER indicator, as well as the number of edits in each individual segment using the TER score. Edits will be analyzed in relation to the processing speed.

4.1. Processing time and processing speed

The processing times (in milliseconds) for all 24 translators were recorded during the assignment but we are presenting here the processing speed (words per minute) per translator. Since we were constrained by the number of fuzzy matches available and the length of the segments, it was not possible with this corpus to have the exact same number of words in all three categories. We had a total of 2,124 words distributed as follows: No match, 749 words, MT match, 757 words and Fuzzy match, 618 words. The data presented in both instances are similar in nature and deal with the same time variable. We thus believe that *Processing speed* is the most relevant to present here.

4.2. Processing speed by Match category

Figure 7 shows the processing speed according to the different match categories for all segments processed in the assignment:

- 1,197 segments processed in the Fuzzy match category: 24 translators processed 50 segments each; except Translators 4 and 9 as explained in section 3.5.
- 1,176 segments processed in the MT match category: 24 translators processed 49 segments each.

- 1,198 segments processed in the No match category) by 24 translators; except Translators 4 and 9.

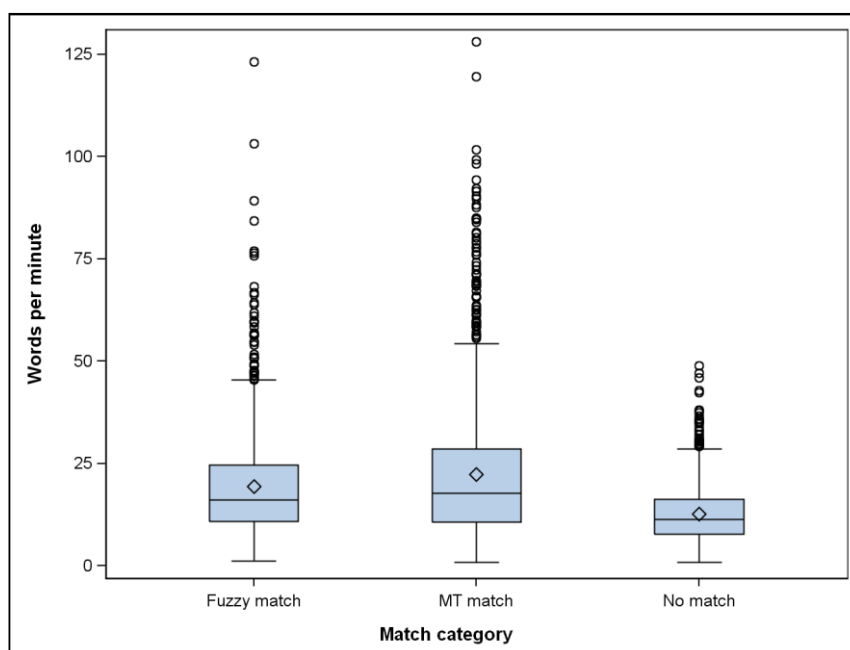


Figure 7: Global processing speed according to Match category

The box-and-whisker diagram shows the distribution of the dataset from 0 to 125 words per minute according to quartiles for the three types of matches. The actual blue box includes data from the first to the third quartile (50 percent of the data). The median value is represented by a black line. The upper whisker represents the data from the third quartile to the maximum value and the lower whisker represents the data from the first quartile to minimum value. The mean value is represented by a diamond and the outliers are presented by circles. Outliers are extreme values that deviate from the rest of the samples.

We can see in Figure 7 that the most words processed per minute on average are in the MT match category, followed by the Fuzzy match, and lastly by the No match category. There are, however, more homogeneous figures for the No match than for the other two categories, where the data have greater dispersion, and there are more outliers. Table 6 shows the descriptive analysis of the data. The Missing column refers to the strings not processed by Translator 4 and 9.

Match category	N	Mean	Median	SD	Min	Max	Missing
Fuzzy match	1197	19.33	16.13	12.67	1.15	123.12	3
MT match	1176	22.34	17.71	17.27	0.89	128.11	0
No match	1198	12.64	11.35	6.97	0.82	48.91	2

Table 6: Descriptive statistics for global processing speed

The mean and median values, as in Figure 7, are higher for MT match, followed very closely by Fuzzy match and then No match, but there is less deviation when processing the latter, showing a more homogeneous processing speed when translators work without a translation proposal. This could be due to the nature of particular segments and to the difference in the proposals from MT and Fuzzy matches. Some segments might require more editing (hence the minimum values 0.15 or 0.89 words per minute), while others might require only reading if the quality of the output is high. For example, an MT match might be a perfect match (we see a maximum value of 128.11 words per minute), a Fuzzy match might require a slight edit (we see a maximum value of 123.3 words per minute), but a No match will always require translators to translate the segment fully, hence processing fewer words per minute in comparison to the other two categories (the maximum value in this category is 48.91 words per minute). This result contrasts with those from our pilot project (Guerberof 2008), where the values showing more dispersion were in the No match category. However, in the pilot the processing speed was calculated per translator (N=translators), while in this case the processing speed is calculated per segment (N=number of segments). Also, the descriptive values per individual translator in this project (see Appendix F) show that all values are higher in all Match categories than those found in the pilot, but the standard deviation is still lower for the No match category. It is difficult to compare results directly because in this case the volume to translate was larger, there were more and different participants, and the engine and corpus were also different, and hence results differ.

We can see that these results show MT increasing translators' speed, in line with previous research (Guerberof 2008, Offersgaard et al. 2008, Masselot and Plitt 2010, De Almeida and O'Brien 2010). Tatsumi (2010) also showed very similar results when studying the English to Japanese language combination: she reported 22.38 words per minute as a mean, and our mean value is very close, at 22.34 words per minute.

In order to test our hypothesis and see if the correlation between MT and Fuzzy matches is significant we applied a linear regression model with repeated measures,

taking as the response variable *logarithm of Words per minute*. For this variable, statistically significant differences are observed ($F=239.96$ and $p < 0.0001$) among the three categories of matches: Fuzzy, MT and No match. From this model, estimates of the mean values of the variable *logarithm of Words per minute* according to the *Match category* have been obtained. To better interpret the results, Table 7 shows the corresponding estimates expressed with the same units as the original variable *Words per minute* according to the Match category and the corresponding confidence intervals of 95 percent (Lower and upper).

Match category	Estimated mean	Lower	Upper
Fuzzy match	15.95	13.79	18.45
MT match	17.21	14.88	19.91
No match	10.79	9.33	12.49

Table 7: Estimated global words per minute

We can observe in Table 7 that the lower and upper estimates overlap in the Fuzzy (13.79 to 18.45 words per minute) and MT match categories (14.88 to 19.91 words per minute), while the same estimates for No match do not (9.33 to 12.49 words per minute). Therefore, we can infer that Fuzzy match and MT match values do not show statistically significant differences while the No match does. In other words, the translators were faster when processing the Fuzzy and MT matches than when translating on their own (No match) and, furthermore, the processing speeds for MT and Fuzzy match segments in the 85-94 percent range are not significantly different.

This is in line with our previous work (Guerberof 2008), where we observed that processing MT matches was faster than processing Fuzzy matches from the 80-90 percent range if the mean value was considered. This followed a study by O'Brien (2006b) that had indicated a correlation between these two categories. Tatsumi (2010) shows a comparison of TM and MT matches in her study (although she uses a low volume of fuzzy matches in comparison to the volume of MT output used) and she concludes that the speed for processing MT matches lies within the fuzzy match range of editing 75 percent TM matches (English to Japanese). However, it needs to be taken into account that when looking at other studies, the language combinations and engines are different and thus it is difficult to directly compare results.

In the commercial sector, JABA Translations, a Portuguese and Spanish language service provider, has reported a correlation similar to the one obtained here (Asia Online 2012) in the English to Portuguese language combination and using a customized Asia Online engine, although we do not currently have any published data on how the

experiment was performed. Autodesk (2011) has published a study on-line carried out in several languages where “MT was roughly as productive as working from matches in the 85-94% category. Note however that this varies significantly across languages”. Our results are not that disparate. This could be an indication of the high correlation level possible between MT matches and high Fuzzy matches if the MT engine is trained with sufficient data, both in terms of high volume (quantity) and high quality data.

Table 8 presents the processing speed (words per minute) per translator (“TR1”, “TR2”, etc.) and match category, with the highest mean and median values highlighted in bold.

Translator	Match category	N	Mean	Median	SD	Min	Max	Missing
TR1	Fuzzy match	50	24.61	22.20	12.68	6.82	61.06	0
	MT match	49	25.88	20.55	16.44	3.40	76.08	0
	No match	50	15.15	14.62	6.09	4.42	29.56	0
TR2	Fuzzy match	50	13.75	12.64	5.85	3.52	29.67	0
	MT match	49	15.26	13.46	7.75	4.61	47.19	0
	No match	50	9.63	9.24	3.43	4.32	22.16	0
TR3	Fuzzy match	50	10.33	8.84	5.46	2.63	29.36	0
	MT match	49	12.31	10.48	8.30	0.89	40.93	0
	No match	50	5.60	5.28	2.74	0.82	13.92	0
TR4	Fuzzy match	49	10.73	9.59	6.09	3.29	35.04	1
	MT match	49	12.17	8.93	8.71	1.36	40.67	0
	No match	49	7.80	7.31	3.24	2.29	15.70	1
TR5	Fuzzy match	50	26.74	22.63	13.31	10.57	66.29	0
	MT match	49	34.57	26.31	24.92	4.88	128.11	0
	No match	50	17.17	16.34	5.81	7.85	36.57	0
TR6	Fuzzy match	50	16.82	14.04	9.71	1.67	45.39	0
	MT match	49	22.33	20.81	14.88	3.19	78.79	0
	No match	50	8.18	7.96	3.81	1.48	19.19	0
TR7	Fuzzy match	50	21.34	18.59	11.08	4.68	51.75	0
	MT match	49	26.28	22.97	18.81	1.04	84.98	0
	No match	50	15.14	13.69	7.50	3.11	37.46	0
TR8	Fuzzy match	50	20.69	19.19	8.84	8.36	42.86	0
	MT match	49	30.13	23.77	25.41	3.51	119.58	0
	No match	50	13.64	12.62	5.14	5.80	35.34	0
TR9	Fuzzy match	48	15.85	12.08	17.61	3.05	123.12	2
	MT match	49	14.07	10.13	11.24	3.55	47.00	0
	No match	49	9.75	9.98	4.60	2.92	19.47	1
TR10	Fuzzy match	50	23.75	23.13	10.61	1.15	49.32	0
	MT match	49	27.55	27.32	14.23	2.36	74.13	0
	No match	50	15.69	15.38	6.68	1.23	34.82	0
TR11	Fuzzy match	50	9.29	9.24	4.62	1.36	19.97	0
	MT match	49	12.07	10.82	6.72	1.75	30.99	0
	No match	50	7.32	7.46	2.74	1.99	14.71	0
TR12	Fuzzy match	50	14.51	12.73	8.74	2.53	43.04	0
	MT match	49	14.44	10.85	12.68	2.20	68.39	0
	No match	50	7.81	7.20	4.17	1.89	18.75	0
TR13	Fuzzy match	50	34.03	29.71	20.29	7.72	89.19	0
	MT match	49	38.58	33.71	22.49	6.48	84.89	0
	No match	50	22.29	19.48	10.37	7.01	48.91	0
TR14	Fuzzy match	50	11.66	9.58	6.79	2.23	31.97	0
	MT match	49	14.39	10.06	14.54	1.34	92.22	0
	No match	50	7.82	7.11	3.87	2.04	20.27	0
TR15	Fuzzy match	50	29.13	25.39	14.11	11.77	76.38	0
	MT match	49	30.81	28.29	17.41	5.58	81.22	0
	No match	50	17.54	17.09	5.52	5.72	35.87	0
TR16	Fuzzy match	50	17.83	14.87	8.78	5.61	48.98	0

Translator	Match category	N	Mean	Median	SD	Min	Max	Missing
TR17	MT match	49	18.78	15.16	11.65	3.85	55.97	0
	No match	50	10.75	10.90	3.19	3.57	18.03	0
	Fuzzy match	50	16.13	14.34	7.40	3.38	41.29	0
TR18	MT match	49	16.74	13.34	11.31	2.90	51.08	0
	No match	50	11.09	10.65	5.94	2.08	29.78	0
	Fuzzy match	50	24.68	21.33	16.74	8.89	103.14	0
TR19	MT match	49	28.88	22.21	18.06	6.20	84.88	0
	No match	50	14.37	13.39	6.73	2.71	31.24	0
	Fuzzy match	50	25.78	23.35	14.99	4.52	63.81	0
TR20	MT match	49	32.91	25.24	23.26	8.39	101.62	0
	No match	50	16.78	15.57	6.23	6.55	34.98	0
	Fuzzy match	50	20.25	19.65	6.20	9.31	36.60	0
TR21	MT match	49	23.03	19.73	13.12	3.56	69.61	0
	No match	50	16.92	15.56	6.78	4.52	42.81	0
	Fuzzy match	50	20.48	17.37	11.97	4.40	66.71	0
TR22	MT match	49	21.24	16.00	14.14	2.11	65.66	0
	No match	50	11.85	11.72	4.82	1.91	24.15	0
	Fuzzy match	50	16.00	14.64	10.42	2.82	43.05	0
TR23	MT match	49	19.55	14.00	17.40	3.26	88.19	0
	No match	50	9.87	8.81	4.95	2.68	25.59	0
	Fuzzy match	50	21.44	19.13	7.93	8.84	43.83	0
TR24	MT match	49	24.72	21.94	9.68	6.64	49.31	0
	No match	50	20.00	19.59	6.22	8.71	33.09	0
	Fuzzy match	50	17.85	14.47	10.02	4.76	56.56	0
TR24	MT match	49	19.54	17.70	11.17	3.68	49.35	0
	No match	50	11.03	11.22	4.48	2.56	24.42	0

Table 8: Descriptive statistics for Words per minute per translator

Table 8 shows that the ranges between minimum and maximum values for each Match category are pronounced, and this means variations in translating segments for each individual translator. We checked to see if the variation in speed was progressive, by looking at the mean values (words per minute) per segment to see if the translators went faster as they progressed in the task, but we saw that the variations were related to each segment and they did not seem to be related to a familiarization with the task (segments were presented to the translators in the same ascendant order). It is also important to remember that there were three initial segments that were processed by translators but were not quantified in the analysis, they were used as a warm up for translators (see section 3.4.3.5). There are only two translators (TR09 and TR12) who were faster when processing Fuzzy matches if the mean value is considered. In the case of TR09, we know she skipped two segments in this category and this could account for the difference. In the case of TR12, she was faster (on average) in Fuzzy matches, even though the maximum speed is found in the MT matches (maximum value). We can also see in Table 8 that, although most translators have higher mean values in the MT match category, the median values are sometimes higher for the Fuzzy match, and the standard deviation for MT match values tends to be higher than in the other two categories. This might of course be because MT segments have a greater probability of being almost

perfect and thus require no change, while the Fuzzy matches used in this project belonging to the 85-94 percent match category would always require some type of interaction on the part of the translator. This can be further seen in some of the maximum values obtained when processing MT. On the other hand, MT segments can also be quite challenging and require substantial work, thus creating the high standard deviation. We will examine this in more depth when we present the TER results in section 4.5.

Having established the correlation between processing MT matches and Fuzzy matches in the 85-94 percent range, we might ask about the similarity between MT matches and a particular match value in that range. For example, were the MT matches post-edited at a speed similar to 85, 87 or 94 percent Fuzzy match? However, we did not have enough segments in each individual match category to draw conclusions on their behavior regarding speed, and then to compare those to the MT matches. For example, we only had three segments in the 85 percent match and two in the 92 percent match. As this was not the initial aim of the study, the segments had not been chosen to analyze this type of correlation. This could be the object of a future study.

4.3. Productivity gain

We want to explore the percentage of gain that translators have when processing MT or Fuzzy matches with respect to their No match processing speed. In order to do this, we model the data using a linear regression model with repeated measures. We present it here using the original variable *Words per minute* according to MT match for better understanding. We have added a column to the right expressing the percentage of gain. This gain has been calculated taking only the estimated mean value (processing speed) of No match and comparing it with the estimated mean value of the other two categories (Fuzzy and MT match). Therefore it does not mean that it is applicable to every single segment processed by translators. We have added as well a Time savings column expressing the percentage of time saved for each translator considering the time invested in translating an amount of words designated as 1 at a No match speed, minus the time invested in doing the same amount of words at a MT or TM speed. We have used the following formula (as in Plitt and Masselot 2010), $x = 1 - \frac{1}{1+y}$ where X is the time saved and y is the productivity gain (Fuzzy and MT match). For example for

Translator 1, the Fuzzy match time savings is: $0.36 = 1 - \frac{1}{1+0.55}$. Table 9 shows all these results.

Translator	Match	Estimated mean	Lower CI 95%	Upper CI 95%	Estimated productivity gain	Time savings
TR1	Fuzzy	21.49	18.36	25.17	55.16%	36%
TR1	MT	21.03	17.93	24.67	51.81%	34%
TR1	NM	13.85	11.83	16.22	.	
TR2	Fuzzy	12.57	11.17	14.14	38.76%	28%
TR2	MT	13.71	12.18	15.44	51.37%	34%
TR2	NM	9.06	8.05	10.19	.	
TR3	Fuzzy	9.11	7.64	10.85	86.49%	46%
TR3	MT	9.61	8.05	11.47	96.89%	49%
TR3	NM	4.88	4.10	5.82	.	
TR4	Fuzzy	9.44	8.08	11.03	32.89%	25%
TR4	MT	9.77	8.36	11.42	37.60%	27%
TR4	NM	7.10	6.08	8.30	.	
TR5	Fuzzy	24.09	20.92	27.74	47.78%	32%
TR5	MT	27.69	24.01	31.93	69.86%	41%
TR5	NM	16.30	14.16	18.77	.	
TR6	Fuzzy	14.21	11.96	16.88	96.48%	49%
TR6	MT	18.07	15.18	21.51	149.9%	60%
TR6	NM	7.23	6.09	8.59	.	
TR7	Fuzzy	18.49	15.47	22.11	38.93%	28%
TR7	MT	20.31	16.96	24.33	52.61%	34%
TR7	NM	13.31	11.13	15.91	.	
TR8	Fuzzy	19.02	16.45	22.00	47.82%	32%
TR8	MT	22.98	19.85	26.62	78.58%	44%
TR8	NM	12.87	11.13	14.88	.	
TR9	Fuzzy	12.23	10.21	14.65	41.24%	29%
TR9	MT	10.81	9.04	12.92	24.75%	20%
TR9	NM	8.66	7.25	10.35	.	
TR10	Fuzzy	20.72	17.59	24.40	47.63%	32%
TR10	MT	23.85	20.22	28.13	69.91%	41%
TR10	NM	14.03	11.92	16.53	.	
TR11	Fuzzy	7.98	6.82	9.33	18.21%	15%
TR11	MT	10.19	8.70	11.93	51.01%	34%
TR11	NM	6.75	5.77	7.89	.	
TR12	Fuzzy	12.12	10.16	14.46	80.13%	44%
TR12	MT	11.19	9.37	13.37	66.34%	40%
TR12	NM	6.73	5.64	8.03	.	
TR13	Fuzzy	28.34	23.98	33.48	40.78%	29%
TR13	MT	31.66	26.75	37.47	57.29%	36%
TR13	NM	20.13	17.03	23.78	.	
TR14	Fuzzy	10.08	8.50	11.96	45.49%	31%
TR14	MT	10.61	8.92	12.61	53.08%	35%
TR14	NM	6.93	5.84	8.22	.	
TR15	Fuzzy	26.25	23.01	29.95	57.78%	37%
TR15	MT	26.31	23.03	30.06	58.11%	37%
TR15	NM	16.64	14.58	18.98	.	
TR16	Fuzzy	16.08	14.14	18.28	56.86%	36%
TR16	MT	15.97	14.03	18.18	55.86%	36%
TR16	NM	10.25	9.01	11.65	.	
TR17	Fuzzy	14.55	12.52	16.90	50.48%	34%
TR17	MT	14.03	12.06	16.33	45.14%	31%
TR17	NM	9.67	8.32	11.23	.	
TR18	Fuzzy	21.11	18.11	24.61	65.02%	39%
TR18	MT	24.31	20.82	28.38	90.05%	47%
TR18	NM	12.79	10.97	14.91	.	
TR19	Fuzzy	21.10	17.94	24.83	34.23%	26%
TR19	MT	26.73	22.68	31.51	70.03%	41%
TR19	NM	15.72	13.36	18.50	.	
TR20	Fuzzy	19.35	17.22	21.75	23.46%	19%

Translator	Match	Estimated mean	Lower CI 95%	Upper CI 95%	Estimated productivity gain	Time savings
TR20	MT	20.19	17.95	22.72	28.81%	22%
TR20	NM	15.67	13.95	17.62	.	
TR21	Fuzzy	17.55	14.86	20.74	63.04%	39%
TR21	MT	16.94	14.31	20.05	57.36%	36%
TR21	NM	10.77	9.11	12.72	.	
TR22	Fuzzy	12.79	10.62	15.41	46.47%	32%
TR22	MT	14.70	12.18	17.74	68.34%	41%
TR22	NM	8.73	7.25	10.52	.	
TR23	Fuzzy	20.16	18.20	22.33	6.21%	6%
TR23	MT	22.89	20.65	25.38	20.62%	17%
TR23	NM	18.98	17.14	21.02	.	
TR24	Fuzzy	15.71	13.56	18.19	56.74%	36%
TR24	MT	16.60	14.31	19.25	65.60%	40%
TR24	NM	10.02	8.65	11.61	.	

Table 9 Estimated WPM and productivity gain per individual translator

All 24 translators have some productivity gain when using MT and Fuzzy matches when estimated mean values are compared. Six translators out of these - 1, 9, 12, 16, 17 and 21 (highlighted in bold in the table) - showed a higher gain when using Fuzzy matches, and consequently 18 translators showed a higher gain when using MT matches.

It is also interesting to note that the productivity gain for MT matches ranges from 20.62 (TR23) to 149.90 percent (TR6), and in the case of Fuzzy matches from 6.21 (TR23) to 96.48 percent (TR6). As we can see, in spite of an increase in both categories, there are different degrees for individual translators. One translator might increase productivity by 150 percent while another might do it by only 20 percent. Further, it can be seen that some translators (for example, translator 9 or 23) might have a higher estimated mean value in MT and Fuzzy, hence the productivity gain, but the confidence intervals show similar values in the three categories, and this might indicate that this gain is not so clear with some translators.

These different degrees can also be seen in the Time savings column: 6 (TR23) to 60 percent (TR6). Prices are determined according to these time savings. In the localization industry, payments for the 85-94 percent range tends to be 60 percent of the full word rate, which would mean 40 percent of time savings for this Fuzzy match range (see Fuzzy match and MT post-editing pricing in Appendix A). Table 9 shows that only TR3, TR6 and TR12 are above 40 percent, and the rest have values below this. Some of these values are close to 40 (for example, TR18 or TR21) but others are quite far from achieving these savings (for example, TR11, TR20 or TR23). The average Time savings for Fuzzy Match for all translators is 32 percent and 37 percent for MT matches. It is important to remember that the time savings in this project are obtained with a high

quality TM and MT output. Therefore, the standard pricing in the industry would appear to be lower than the one that should be applied to this particular set of fuzzy matches, and as a consequence to the MT matches. However, in our project, the translators are not working with the original tool (SDL Trados 2007) where changes to be made in each fuzzy match are highlighted. Therefore, it is possible that the time savings using the original tool are higher than using our tool.

Figure 8 shows the productivity gain for all translators if an estimated mean value of speed (words per minute) is assumed. We are using a continuous line to connect all translators so the relation can be seen more clearly; however, the actual estimated mean is a dot between the two axes for each translator.

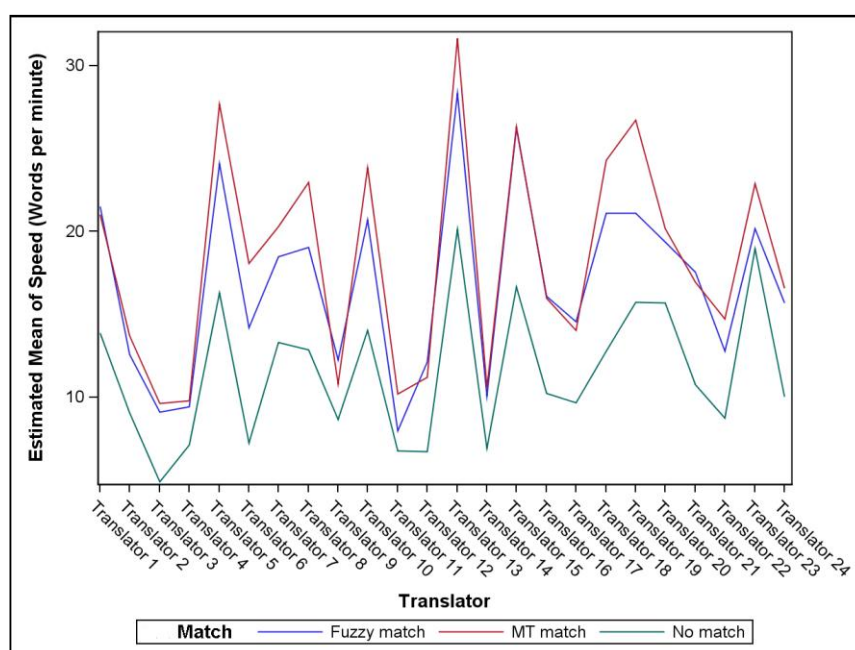


Figure 8 Estimated mean of speed per translator

The green line, representing the No match category, is the lower line for all translators, the blue line (Fuzzy match) is close to the red line (MT match) but it is higher for the six translators mentioned above, and the red line is higher for the majority of translators.

TR13 obtained the highest processing speed with MT - 31.66 words per minute - and the same translator shows the second highest processing speed with TM - 28.33 words per minute. TR03 shows the lowest estimated processing speed with No match, 4.88 words per minute. As we saw in our previous project (Guerberof 2008), there is great disparity between translators when it comes to productivity, making it difficult to use standard measurements and hence to set a correct pricing that would compensate

everyone’s real effort. If we consider the 2,500 words per day metric and this volume is paid per word, a slower translator will get paid less per day and a faster translator will be paid more per day, but the amount paid for the full number of words (the job) is still the same. In this sense, we can say that the payment system is “fair” if the job is considered. However, if a percentage of the word edit or post-edit is paid, for example 50 percent, assuming an average productivity increase for all translators alike (for the same job), then a translator that has a lower productivity increase when using these proposals than when translating on his or her own might benefit less than a translator that has a higher productivity increase when using these same proposals, and therefore this originally fast translator might be paid less for the same job. The situation is similar with translation memories and machine translation, as seen above, and it might be difficult to find a satisfactory solution to determine a “fair” price. Nevertheless, we would like to explore the relationship between speed and productivity increase in section 4.4.

4.3.1. Estimated words per day

Table 10 shows the mean estimated speed (words per minute) from our model, (Table 9), per translator extrapolated to eight hours of work (the standard used in the industry to calculate the productivity in a project) according to the match category.

Translator	Match category	Estimated mean	Estimated words day
TR1	Fuzzy match	21.49	10,315
TR1	MT match	21.03	10,094
TR1	No match	13.85	6,648
TR2	Fuzzy match	12.57	6,034
TR2	MT match	13.71	6,581
TR2	No match	9.06	4,349
TR3	Fuzzy match	9.11	4,373
TR3	MT match	9.61	4,613
TR3	No match	4.88	2,342
TR4	Fuzzy match	9.44	4,531
TR4	MT match	9.77	4,690
TR4	No match	7.1	3,408
TR5	Fuzzy match	24.09	11,563
TR5	MT match	27.69	13,291
TR5	No match	16.3	7,824
TR6	Fuzzy match	14.21	6,821
TR6	MT match	18.07	8,674

Translator	Match category	Estimated mean	Estimated words day
TR6	No match	7.23	3,470
TR7	Fuzzy match	18.49	8,875
TR7	MT match	20.31	9,749
TR7	No match	13.31	6,389
TR8	Fuzzy match	19.02	9,130
TR8	MT match	22.98	11,030
TR8	No match	12.87	6,178
TR9	Fuzzy match	12.23	5,870
TR9	MT match	10.81	5,189
TR9	No match	8.66	4,157
TR10	Fuzzy match	20.72	9,946
TR10	MT match	23.85	11,448
TR10	No match	14.03	6,734
TR11	Fuzzy match	7.98	3,830
TR11	MT match	10.19	4,891
TR11	No match	6.75	3,240
TR12	Fuzzy match	12.12	5,818
TR12	MT match	11.19	5,371
TR12	No match	6.73	3,230
TR13	Fuzzy match	28.34	13,603
TR13	MT match	31.66	15,197
TR13	No match	20.13	9,662
TR14	Fuzzy match	10.08	4,838
TR14	MT match	10.61	5,093
TR14	No match	6.93	3,326
TR15	Fuzzy match	26.25	12,600
TR15	MT match	26.31	12,629
TR15	No match	16.64	7,987
TR16	Fuzzy match	16.08	7,718
TR16	MT match	15.97	7,666
TR16	No match	10.25	4,920
TR17	Fuzzy match	14.55	6,984
TR17	MT match	14.03	6,734
TR17	No match	9.67	4,642
TR18	Fuzzy match	21.11	10,133
TR18	MT match	24.31	11,669
TR18	No match	12.79	6,139
TR19	Fuzzy match	21.1	10,128
TR19	MT match	26.73	12,830
TR19	No match	15.72	7,546
TR20	Fuzzy match	19.35	9,288
TR20	MT match	20.19	9,691
TR20	No match	15.67	7,522
TR21	Fuzzy match	17.55	8,424
TR21	MT match	16.94	8,131

Translator	Match category	Estimated mean	Estimated words day
TR21	No match	10.77	5,170
TR22	Fuzzy match	12.79	6,139
TR22	MT match	14.7	7,056
TR22	No match	8.73	4,190
TR23	Fuzzy match	20.16	9,677
TR23	MT match	22.89	10,987
TR23	No match	18.98	9,110
TR24	Fuzzy match	15.71	7,541
TR24	MT match	16.6	7,968
TR24	No match	10.02	4,810

Table 10: Estimated words per day

As was explained in section 3.3, extrapolating results from a 3 to 4 hour assignment to 8 hours will not reflect the “real” productivity of a translator in an 8-hour day (as O’Brien 2011 also remarks), as we can infer from the high processing speeds in the No match categories shown in the table, for example. However, this table gives perhaps a clearer indication of the productivity increase experienced by this group, how similar the number of words processed for Fuzzy (85-94 percent) and MT match are for all 24 translators, and the variability in speed among them.

4.4. Speed groups and processing speeds

Our sub-hypothesis says that translators with higher processing speeds, measured in words per minute, when translating the No match segments will show less productivity gain when post-editing the proposed text from MT or TM than the translators with lower processing speeds when working with the same set of segments. We had observed this pattern in our previous project (Guerberof 2008), where faster translators would not necessarily increase their productivity in spite of using MT or TM segments as proposals. Therefore, we had to define a process to measure this gain with respect to the No match segments, and then verify if the translators with had a higher speed would again show less productivity gain when using MT or Fuzzy match segments.

4.4.1. Standardized Words per Minute per translator

We posit that a translator has an intrinsic speed (the speed in No match segments without a translation aid) and we define that variable as “*Standardized Words per minute by translator with respect to No Match*” (Std_WM). This is thus a

standardization of the variable *Words per minute* with respect to the mean value and the standard deviation in the No match segments.

Per translator we calculated:

- meanNM = mean value of No match segments;
- stdNM = standard deviation of No match segments for each translator;
- std_WM= Standardized Words per minute by translator with respect to No Match.

Finally, for the Fuzzy match and MT match segments a variable of interest is calculated:

- $std_WM = (\text{Words_per_minute [in the Fuzzy or MT match category]} - \text{meanNM})/\text{stdNM}$

The coefficient will be 0 if there are no gains with respect to the No match, and it will be positive if there are gains in speed. If there are segments translated at a lower speed with respect to the No match, the value will be negative.

For example: a translator processes 5 No match segments with the following speeds: 13, 9, 8, 8 and 12. The mean value is 10 and the standard deviation is 2.35. The same translator processes Fuzzy match or MT match segments with a speed of 15 words per minute (higher average speed than with No match). Then, we have $std_WM = 2.13$. The value is positive and we can therefore conclude that this translator had a gain in speed. Below we see some samples for individual translators. Figure 9 shows the values for Translator 3: the Fuzzy match (blue line) and MT match (red line) speeds compared to the mean reference value for No match. Translator 3 had the lowest processing speed in No match as we saw in section 4.1, the mean value for No match was 5.60 words per minute. In Figure 10, we see the *standardized* speed for Translator 3 with respect to No match where the reference line is 0.

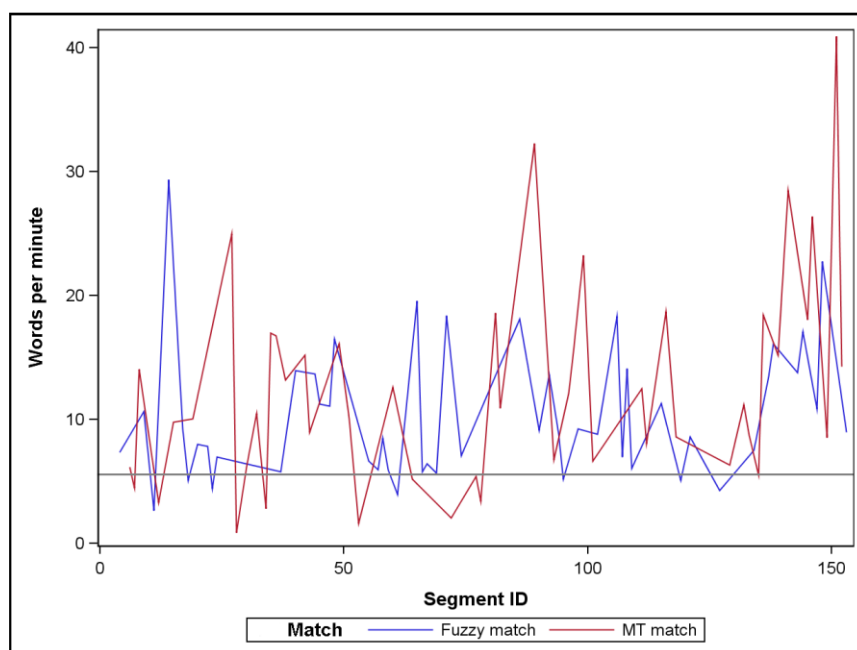


Figure 9: Sample for TR3 WPM

The blue and red lines indicate the processing speed of the Fuzzy match and MT match categories respectively.

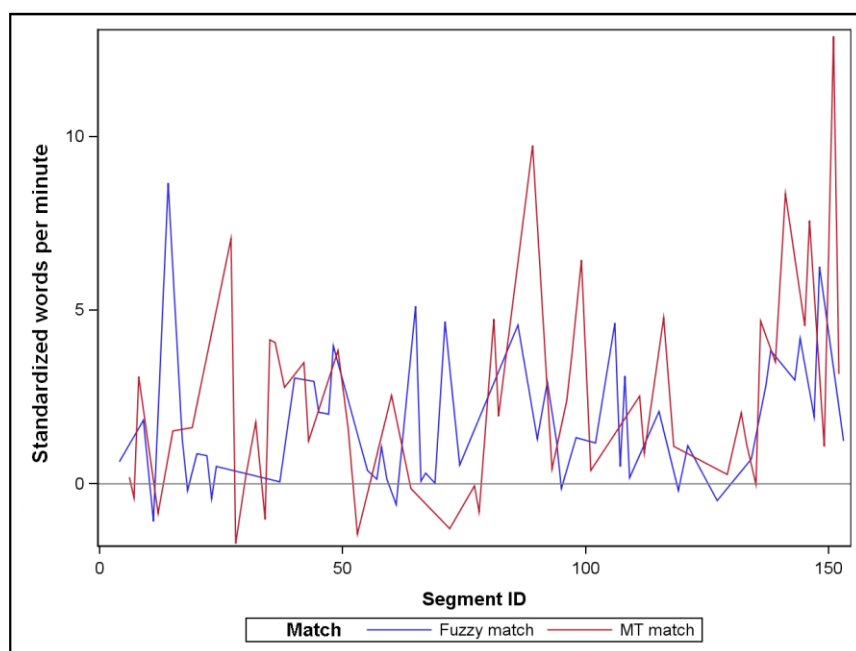


Figure 10: Sample for TR3 Standardized WPM

Figure 10 shows the standardized speed for Translator 3 with the base line for No match set as 0. Each segment that was translated faster would be above 0 and those that were translated slower would be below 0. Here we can see that many more segments are above 0, thus indicating that this translator made favorable use of the MT and Fuzzy

matches. On the other extreme, Translator 13 was the fastest translator in the No match category, at 22.29 words per minute.

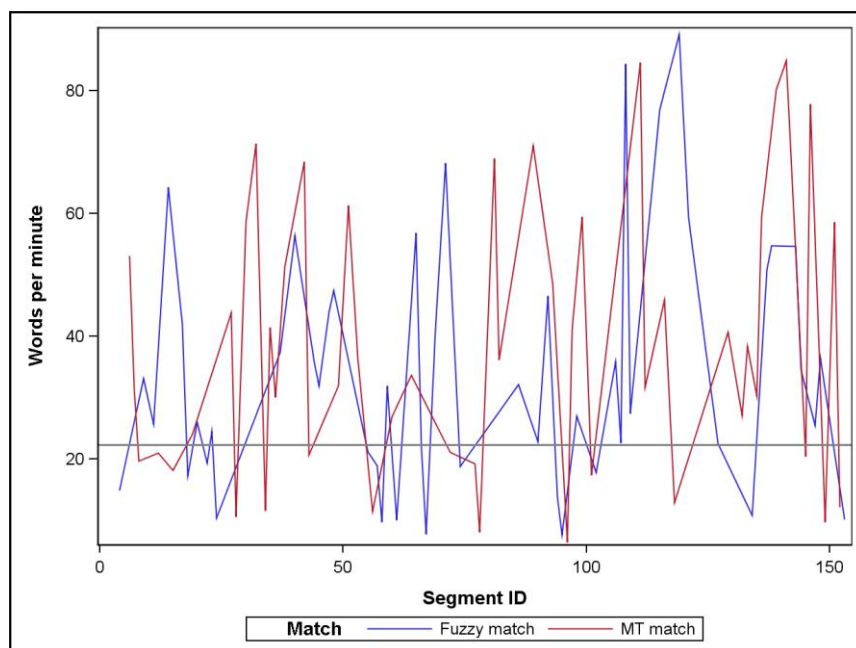


Figure 11: Sample for TR13 WPM

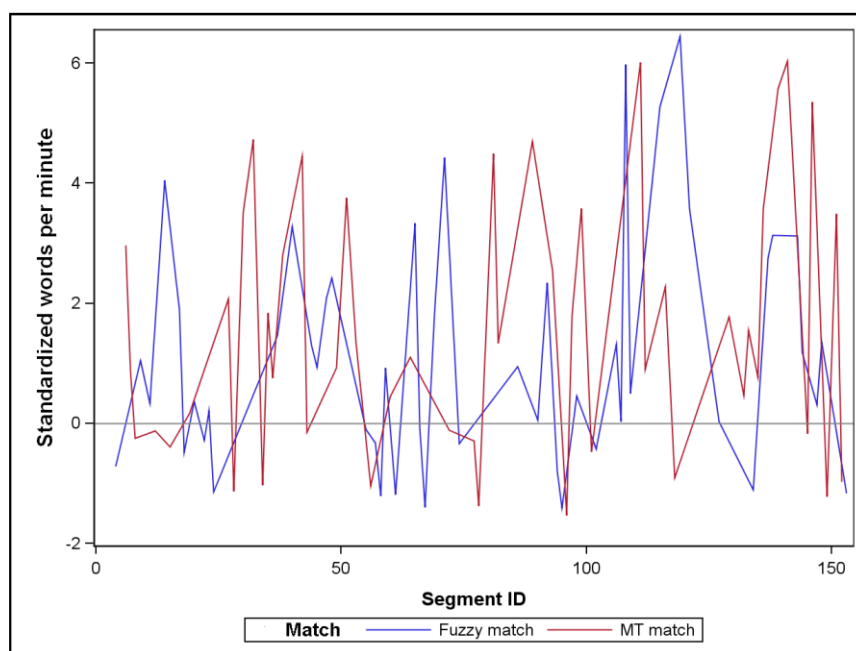


Figure 12: Sample for TR13 Standardized WPM

The standardized words per minute for Translator 13 shows a similar pattern but the “benefit” is lower than in the case of Translator 3. We can see more segments below 0, meaning that translator 13 avails less of the MT and Fuzzy matches than Translator 3, as our sub-hypothesis claims.

Now, in order to test this sub-hypothesis, we need to group translators according to their speed when translating “by themselves”. We created three No Match speed groups by taking the mean value of translator’s word per minute in the No match category and thus defining the following:

- Group 1: Less than 10 words per minute. Translators: 2, 3, 4, 6, 9, 11, 12, 14 and 22.
- Group 2: From 10 (included) to 15 words per minute. Translators: 8, 16, 17, 18, 21 and 24.
- Group 3: 15 or more words per minute. Translators: 1, 5, 7, 10, 13, 15, 19, 20 and 23.

Table 11 provides the descriptive statistical data for the Standardized words per minute according to the speed groups. We have classified the data, taking all the segments per each category belonging to that particular group of translators (this is represented by the N value in Table 11).

Group	Match	N	Mean	Median	SD	Min	Max	Missing
Group 1:	Fuzzy match	447	1.33	0.83	2.28	-2.18	24.63	3
	MT match	441	1.90	0.97	3.13	-2.03	21.79	0
Group 2:	Fuzzy match	300	1.55	0.91	2.24	-1.61	13.20	0
	MT match	294	2.11	1.07	3.29	-2.17	20.60	0
Group 3:	Fuzzy match	450	1.18	0.72	1.97	-2.18	10.67	0
	MT match	441	1.80	1.10	2.88	-2.17	19.09	0

Table 11: Standardized WPM with respect to No match per group

Group 1 shows positive mean values with respect to No match; Group 2 also shows positive values that are slightly higher than those of Group 1 (therefore, Group 2 took more advantage, in terms of speed, of MT and Fuzzy matches than Group 1); and finally Group 3 shows a positive value but lower than that of Group 1 (therefore Group 3 took less advantage than Group 1 and 2). At first glance, it might appear that Group 2, with speeds from 10 to 15 words per minute, is the group that takes the most advantage of the MT and Fuzzy matches. However, the range values are wide and the standard deviations are correspondingly high.

We present the results of the linear model with repeated measures taking *Standardized Words per minute with respect to No match* as the response variable and *Match* and *Speed group* as explanatory factors. We find statistically significant differences among Match categories (F=9.30; p=0.0023) but there are no statistically

significant differences between the Speed groups, nor in the interaction between the two Match categories and Speed groups.

In order to see this clearly, from the model we calculate the estimated mean of the *Standardized Words per minute with respect to No match* according to Speed group and Match category. We show below the value for this estimation and the corresponding 95 percent confidence intervals.

Match category	Estimated mean	Lower 95%	Upper 95%
Fuzzy match	1.35	1.09	1.62
MT match	1.94	1.67	2.20

Table 12: Estimated standardized Words per minute

Here we can see that there are statistically significant differences between Fuzzy match and MT match once the values have been standardized with respect to the No match. In this case, unlike the previous results we obtained in our first hypothesis (see 4.2), we are comparing standardized speeds with respect to No match, hence the difference in the results. It is interesting to see that if the intrinsic translators' speed is considered (No match), there are statistically significant differences between the MT match and Fuzzy match and that the processing speed for MT is significantly higher. Let us return to the Speed groups again.

Speed group	Mean	Lower 95%	Upper 95%
Group 1	1.61	1.31	1.92
Group 2	1.83	1.46	2.20
Group 3	1.49	1.19	1.79

Table 13: SWPM and speed groups

As we can see in Table 13, Group 2 has the highest mean value, followed by Group 1 and lastly Group 3. However, the confidence intervals show that there are no statistically significant differences to support the sub-hypothesis. We could say, in view of this, that the benefits when using MT or TM in terms of the standardized words per minute with respect to No match are not affected significantly by the speed of translators in No match. In very simple and plain words, no matter whether you are a slow or fast translator you will be equally benefited by MT or Fuzzy matches. Nevertheless, we can observe that for Group 3, the variable drops and the benefits are lower, but not significantly so.

Table 14 shows the interaction between both variables: Match and Speed groups.

Match	Speed group	Estimated	Lower 95%	Upper 95%
Fuzzy match	Group 1	1.33	0.91	1.76
Fuzzy match	Group 2	1.55	1.03	2.07
Fuzzy match	Group 3	1.18	0.76	1.61
MT match	Group 1	1.90	1.47	2.32
MT match	Group 2	2.11	1.59	2.64
MT match	Group 3	1.80	1.38	2.23

Table 14: SWPM and speed groups & Match categories

If we consider only the Fuzzy match (rows 1, 2 and 3), we do not see any significant differences within the confidence intervals in the three groups, although Group 3 drops (row 3) again in the estimated standardized value. If we consider only the MT match (rows 4, 5 and 6), we observe the same pattern: no significant differences but Group 3 drops in value again. We can also see that Group 2 seems to be the one that benefits most from both MT and Fuzzy matches, but this is not statistically significant. Moreover, the estimated mean values for MT matches are higher for the three speed groups (as we saw above) than the estimated mean values for Fuzzy matches. Although no statistically significant differences are observed, the fact that we see certain differences among the groups might constitute an interesting aspect to study further.

4.5. Speed and number of edits: TER

We would like to turn now to the number of edits that translators made in the segments, to observe if depending on the type of proposal, Fuzzy or MT, the translators implemented different numbers of edits. We know now that the type of proposal did not affect their overall speed but did it affect the number of changes made? What is the relationship between number of changes and processing speed? Other studies, such as Tatsumi (2009), Tatsumi and Roturier (2010) and O'Brien (2011), have explored the relationship between automatic metrics and processing speed in post-editing. Tatsumi and Roturier examined correlations between the General Text Matcher metric (GTM) (Melamed et al. 2003, Turian et al. 2003) and processing speed. They found that post-editors made either a similar number of edits at very different speeds, or that these scores assessed the technical effort (number of edits) more accurately than the temporal effort (words per minute). O'Brien also explored GTM as a predictor of post-editing productivity but more interestingly she examined the relation between post-task TER scores and productivity (as we have done for the present study) and found a correlation between increase in the TER score, on average, and decrease in processing

speed, although she pointed out that this could not necessarily be applied to individual segments.

When looking at the wide ranges in processing speed for MT, we have also established the possibility that some of these segments might have been perfect matches that required no change while others required substantial work. It would be interesting to test if this was the case for all translators. With these ideas in mind, we calculated the TER score per translator for all Fuzzy and MT match segments separately. As we explained earlier, a TER value of 0 means that no changes were applied to the proposal: the higher the TER, the higher the number of changes made.

4.5.1. Correlation between TER, Levenshtein, and Olivier and Hand

Before looking at the TER scores (see section 3.4.3.10), we explore the correlation between TER and Olivier and Hand (Olivier and Hand 1996), which calculates the similarity between two strings. Olivier and Hand gives a score of 100 if the two strings are identical and 0 if they are completely different. This score was automatically calculated by the CrossLang post-editing tool and therefore it was interesting to compare these “automatic” values with those of TER. We also explore the correlation between TER and Levenshtein distance. Levenshtein looks at the minimum number of edits needed to transform one string into another (per character) and it gives a score. This distance was calculated automatically by the SAS software and therefore it was interesting to compare these “automatic” metric with those of TER to see if our findings using TER would be similar if using the other two metrics.

There is a negative Pearson correlation (-0.83175 $p < .0001$) between Olivier and Hand, and TER and a positive Pearson correlation (-0.62923 $p < .0001$) between Levenshtein and TER. Therefore, the correlation between Olivier and Hand, and TER is slightly stronger. In Figure 13 and Figure 14 we can see that when TER is closer to 0 there is more correlation with the other two indexes (and this is understandable as these are the strings where no changes are made) and as TER becomes higher there are more differences with the other two indexes.

In view of these data, we can say that the results obtained using Olivier and Hand, and Levenshtein would be in general similar to the ones we will present using TER but not equal, and that differences will be higher when more changes are made in each segment.

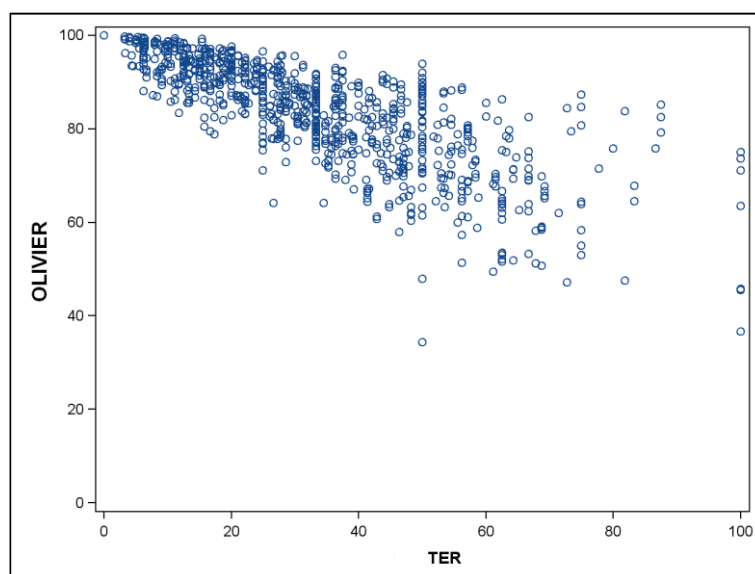


Figure 13: Correlation between TER, and Olivier and Hand

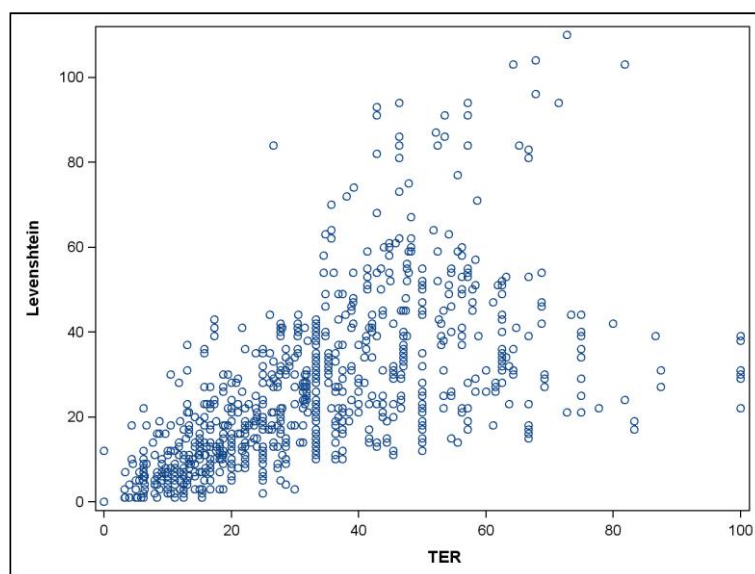


Figure 14: Correlation between TER and Levenshtein

4.5.2. TER indicator: segments edited

Having described this correlation, we return now to the TER scores. Firstly, we created a TER indicator to explore the number of segments where translators had made edits.

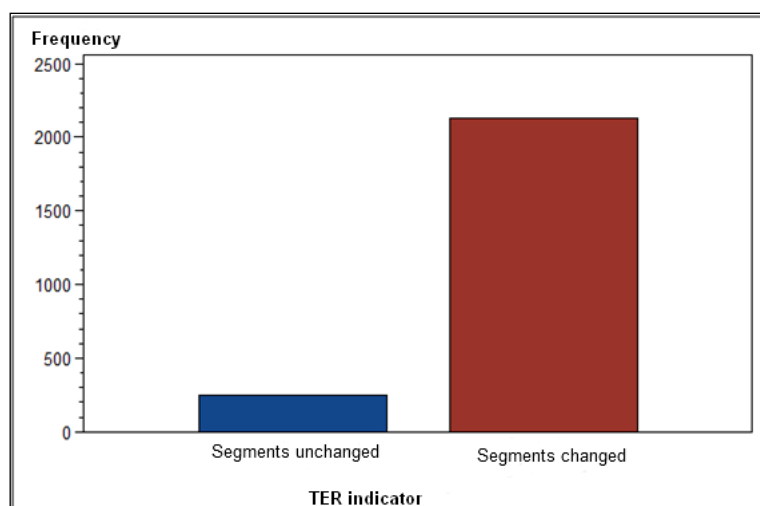


Figure 15: TER indicator for all Fuzzy and MT matches

Figure 15 graphically shows that 89.65 percent of the segments were edited and 10.35 percent were not. And we can see below that this proportion of changed versus unchanged segments is similar across translators.

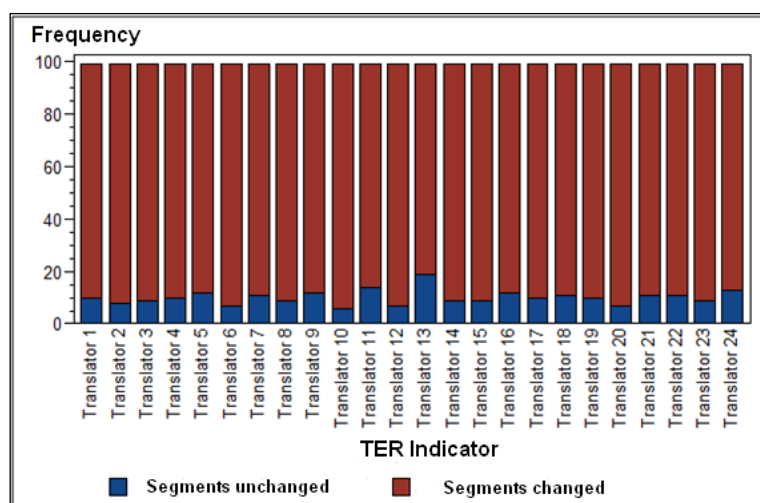


Figure 16: TER indicator per translator

From Figure 16 we can see that Translator 13, for example, who was the fastest among all 24 translators when translating No match, is the one that left the most segments unchanged (19 segments) so this could indicate a relationship between speed and unchanged segments. Table 15 shows the same data in a table format. The data are sorted according to percentage of segments changed starting from the maximum value.

Translator	Segments unchanged		Segments changed		Total
	N	Row %	N	Row %	
Translator 10	6	6.06	93	93.94	99
Translator 6	7	7.07	92	92.93	99

Translator	Segments unchanged		Segments changed		Total
	N	Row %	N	Row %	
Translator 12	7	7.07	92	92.93	99
Translator 20	7	7.07	92	92.93	99
Translator 2	8	8.08	91	91.92	99
Translator 3	9	9.09	90	90.91	99
Translator 8	9	9.09	90	90.91	99
Translator 14	9	9.09	90	90.91	99
Translator 15	9	9.09	90	90.91	99
Translator 23	9	9.09	90	90.91	99
Translator 1	10	10.1	89	89.9	99
Translator 4	10	10.1	89	89.9	99
Translator 17	10	10.1	89	89.9	99
Translator 19	10	10.1	89	89.9	99
Translator 7	11	11.11	88	88.89	99
Translator 18	11	11.11	88	88.89	99
Translator 21	11	11.11	88	88.89	99
Translator 22	11	11.11	88	88.89	99
Translator 5	12	12.12	87	87.88	99
Translator 9	12	12.12	87	87.88	99
Translator 16	12	12.12	87	87.88	99
Translator 24	13	13.13	86	86.87	99
Translator 11	14	14.14	85	85.86	99
Translator 13	19	19.19	80	80.81	99

Table 15: TER indicator per translator

Figure 17 shows this same percentage of TER but in relation to the proposed text, Fuzzy or MT matches.

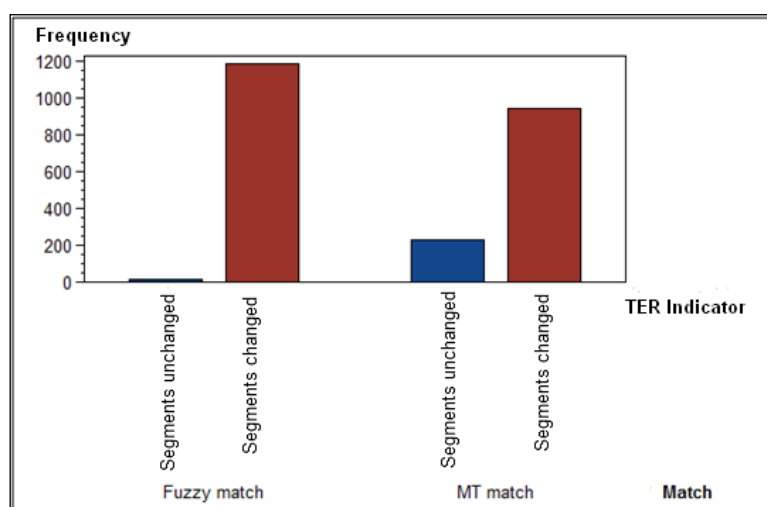


Figure 17: TER indicator for Fuzzy and MT matches

The contrast here is greater: 19.56 percent of MT match segments were not changed, as opposed to 1.33 percent in the Fuzzy match category. There are statistically significant differences between the proportion of segments unchanged in MT matches compared to the ones in Fuzzy matches ($F=121.49$ and $p<.0001$).

The fact that segments were left unchanged indicates that all translators found that some MT segments did not require any change while Fuzzy match segments did, and this is a clear indication that the quality of the output for MT was high, as we had

seen initially (section 3.4.3.4). This appears to be the experience for all translators. We start seeing here that the fact that this output had a percentage of almost perfect matches (in the range of 20 percent if judged by the figures from Figure 17) affects the productivity of translators and produces a processing speed comparable statistically to that of processing a high fuzzy match in a translation memory.

4.5.3. TER score: edits per segment

The TER indicator tells us the proportion of segments that were changed. However, it is not telling us the extent of these edits per segment. It could be that some MT segments were not changed, but those that were changed needed substantial work. To examine this, we need to turn to the TER score.

Let us now analyze the TER scores according to the match category.

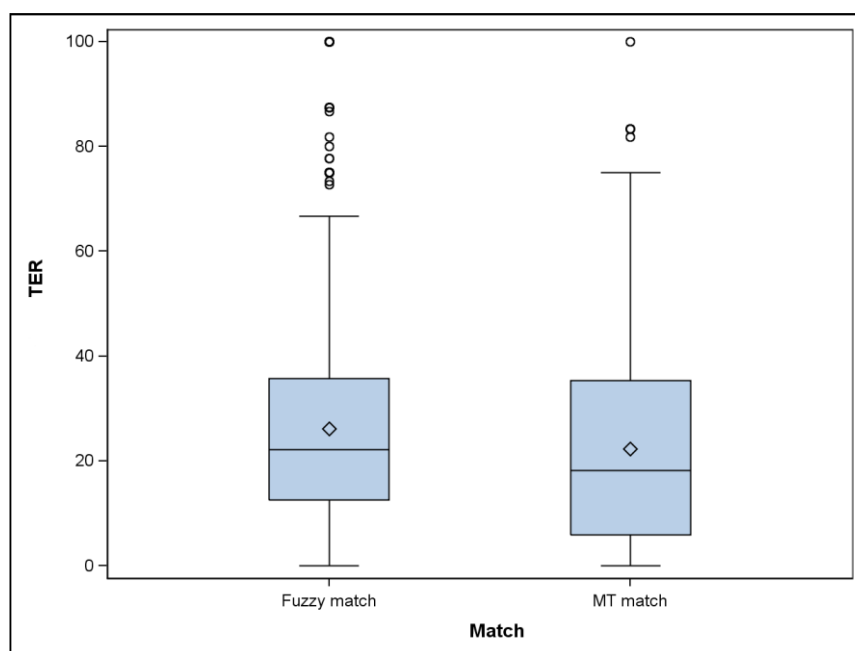


Figure 18: TER score for Fuzzy and MT matches

The mean and median values are not dissimilar between the two categories but the MT match presents more values closer to 0, thus indicating, as we pointed out previously, that translators found almost perfect matches in this category.

Table 16 shows the descriptive values for TER.

Match	N	Mean	Median	SD	Min	Max	Missing
Fuzzy match	1197	26.13	22.22	17.89	0.00	100.00	3
MT match	1176	22.27	18.18	19.96	0.00	100.00	0

Match	N	Min	Q1	Median	Q3	Max	Range	Q Range
Fuzzy match	1197	0.00	12.50	22.22	35.71	100.00	100.00	23.21
MT match	1176	0.00	5.88	18.18	35.29	100.00	100.00	29.41

Table 16: Descriptive values for TER

Fuzzy match segments have higher mean and median values. The minimum and maximum values are 0 and 100 respectively in both categories and this means that there were segments where there were no edits (0) and others that were completely changed (100) in both categories. The high standard deviations indicate that at segment level there were different numbers of edits and that this deviation was higher in MT matches, but this is understandable: as we saw earlier, there were segments left unchanged (0 value) and others completely changed (100 value). A linear regression model with repeated measures was applied taking *TER* as response variable and *Match category* as explanatory variable. For the *TER* variable there were statistically significant differences between the Match categories ($F=24.80$; $p<0.0001$). From this model, estimated mean values were obtained for the variable *TER according to Match category*. Table 17 illustrates this data.

Effect	Match	Estimated mean	Standard Error	Lower	Upper
Match	Fuzzy match	26.14	0.59	24.97	27.30
Match	MT match	22.27	0.60	21.10	23.44

Table 17: Estimated mean values for TER

We observed that the estimated mean for TER in Fuzzy match segments is between 24.97 and 27.30 (lower and upper confidence intervals), and for MT match segments between 21.10 and 23.44. The mean for TER in Fuzzy matches is higher and this is a statistically significant difference. This indicates that although segments were processed with similar processing speeds in Fuzzy and MT match categories (as we saw in the processing speed section), the number of edits was statistically different in these two categories with more changes being made in the Fuzzy match category. There is a statistical difference between the two categories but this difference is relatively low (4 points out of 100). We then verified if this was the case for each individual translator in the project, that is, rather than looking per group of segments (observations) we looked at the values for Fuzzy and MT match per translator. There were significant differences only for Translator 13 and Translator 22 ($F=24.80$; $p<0.0001$). Figure 19 illustrates these findings.

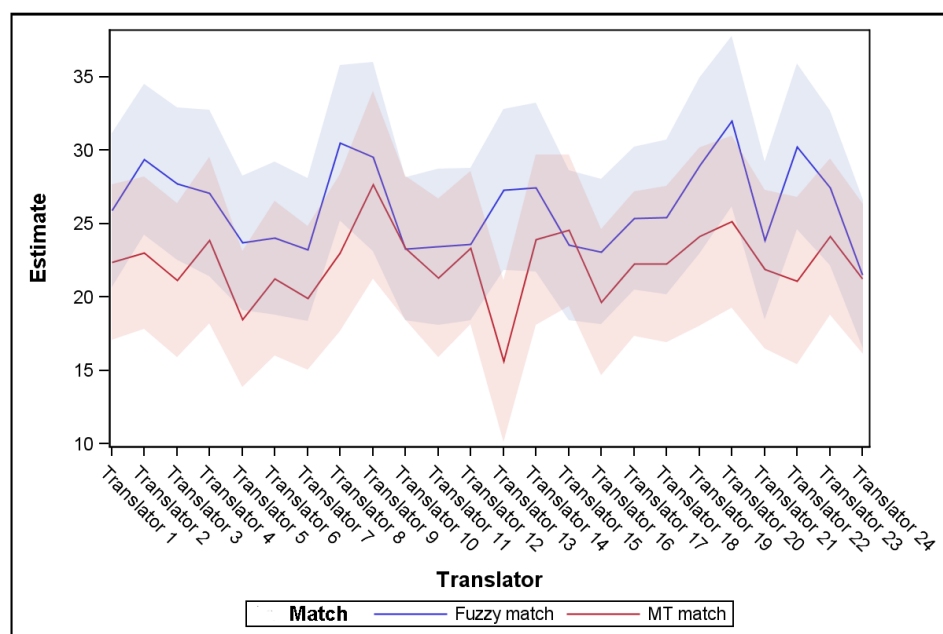


Figure 19: Estimated TER per Translator.

Translators 13 and 22 made fewer edits in MT match than in Fuzzy match and there is a statistically significant difference between both categories. However, the other translators show lower values in MT match than in Fuzzy match (despite Translators 15 having a higher value in MT and Translators 10 and 24 showing very similar values in both categories) but this difference is not statistically significant. In conclusion, if we look at the global model (segment level) with 2,373 observations (all the MT and Fuzzy match segments processed by all translators) there are statistically significant differences, even though per translator (with fewer observations) this is only the case for these two translators.

Table 18 shows the descriptive values for TER scores per segment sorted per ascending mean.

Segment ID	N	Mean	Median	SD	Min	Max	Missing	Match category
42	24	0.52	0	1.76	0	6.25	0	MT
8	24	1.94	0	5.55	0	26.67	0	MT
99	24	1.96	0	5.39	0	17.65	0	MT
141	24	2.08	0	4.43	0	16.67	0	MT
35	24	2.5	0	3.69	0	12	0	MT
81	24	2.6	0	7.35	0	25	0	MT
133	24	2.78	0	4.91	0	11.11	0	MT
27	24	3.29	5.26	2.6	0	5.26	0	MT
96	24	3.75	0	9.24	0	30	0	MT
151	24	4.38	5	2.68	0	10	0	MT
116	24	4.51	4.17	4.9	0	16.67	0	MT
89	24	5.15	5.88	6.79	0	29.41	0	MT
139	24	6.72	6.45	4.55	0	16.13	0	MT
48	24	8.61	6.67	3.67	6.67	20	0	Fuzzy
86	24	9.09	9.09	0	9.09	9.09	0	Fuzzy

Segment ID	N	Mean	Median	SD	Min	Max	Missing	Match category
72	24	9.11	6.25	8.64	0	31.25	0	MT
106	23	9.57	10	2.09	0	10	1	Fuzzy
119	24	9.64	6.25	6.89	6.25	37.5	0	Fuzzy
9	24	10.61	9.09	5.13	9.09	27.27	0	Fuzzy
146	24	10.75	10.53	3.95	0	26.32	0	MT
65	24	11.11	11.11	0	11.11	11.11	0	Fuzzy
108	24	11.46	12.5	3.98	6.25	25	0	Fuzzy
45	23	11.59	8.33	8.97	0	33.33	1	Fuzzy
17	24	11.74	9.09	5	9.09	27.27	0	Fuzzy
148	24	11.88	10	3.85	10	25	0	Fuzzy
152	24	12.5	8.33	7.77	0	25	0	MT
111	24	12.88	13.64	5.47	0	22.73	0	MT
18	24	13.02	12.5	2.55	12.5	25	0	Fuzzy
92	24	13.33	13.33	0	13.33	13.33	0	Fuzzy
6	24	13.5	12	6.44	8	32	0	MT
14	24	13.77	13.04	2.77	8.7	21.74	0	Fuzzy
149	24	14.06	12.5	8.5	0	37.5	0	MT
136	24	14.38	12.5	5.95	10	30	0	MT
47	24	14.58	14.29	1.46	14.29	21.43	0	Fuzzy
24	24	14.81	11.11	7.08	0	33.33	0	Fuzzy
93	24	15.12	18.52	6.99	0	25.93	0	MT
30	24	15.22	13.04	12.16	0	47.83	0	MT
58	24	15.48	14.29	4.03	14.29	28.57	0	Fuzzy
82	24	15.48	14.29	4.03	14.29	28.57	0	MT
55	24	16.41	12.5	11.63	0	43.75	0	Fuzzy
115	24	16.67	15	7.76	0	30	0	Fuzzy
38	24	16.88	17.5	11.96	5	45	0	MT
61	24	17.32	15.79	8.28	10.53	36.84	0	Fuzzy
71	24	17.4	17.65	4.42	0	23.53	0	Fuzzy
51	24	17.76	15.79	13.33	0	63.16	0	MT
143	24	18.06	13.33	7.48	13.33	40	0	Fuzzy
147	24	18.18	18.18	6.57	9.09	36.36	0	Fuzzy
101	24	18.29	19.44	7.41	0	33.33	0	MT
11	24	19.87	15.39	13.01	0	38.46	0	Fuzzy
40	24	20	13.33	12.28	6.67	53.33	0	Fuzzy
97	24	20.04	19.05	5.06	14.29	33.33	0	MT
7	24	20.83	25	9.52	0	25	0	MT
49	24	21.17	20	9.4	12	44	0	MT
132	24	21.35	18.75	7.12	6.25	43.75	0	MT
37	24	21.99	16.67	7.04	16.67	38.89	0	Fuzzy
121	24	22.14	25	9.21	6.25	37.5	0	Fuzzy
59	24	22.16	29.55	16.65	0	54.55	0	Fuzzy
145	24	22.81	21.05	6.14	5.26	36.84	0	MT
60	24	23.03	20.83	10.25	8.33	55.56	0	MT
64	24	23.61	25	4.71	8.33	33.33	0	MT
109	24	24.77	22.22	14.28	5.56	61.11	0	Fuzzy
94	24	25	22.22	6.75	22.22	44.44	0	Fuzzy
20	24	25.52	25	2.55	25	37.5	0	Fuzzy
144	24	27.31	22.22	8.66	22.22	44.44	0	Fuzzy
12	24	28.41	27.27	3.07	27.27	36.36	0	MT
138	24	29.17	27.27	6.55	27.27	54.55	0	Fuzzy
53	24	31.44	36.36	16.08	0	63.64	0	MT
57	24	32.07	30.44	8.29	8.7	43.48	0	Fuzzy
56	24	32.18	30.56	6.13	22.22	50	0	MT
22	24	32.43	28.26	11.89	17.39	65.22	0	Fuzzy
107	24	32.78	33.33	15.09	0	66.67	0	Fuzzy
44	23	33.85	28.57	9.44	28.57	64.29	1	Fuzzy
134	24	34.52	28.57	11.85	28.57	57.14	0	Fuzzy
153	24	36.11	33.33	13.38	16.67	58.33	0	Fuzzy
32	24	36.57	44.44	18.01	0	61.11	0	MT
127	24	36.63	33.33	14.69	16.67	58.33	0	Fuzzy
112	24	38.38	39.47	14.6	10.53	63.16	0	MT
43	24	40.28	40	4.16	26.67	53.33	0	MT
15	24	40.44	41.18	8.55	11.77	52.94	0	MT
69	24	40.63	41.67	13.53	0	75	0	Fuzzy
98	24	42.05	36.36	13.08	27.27	81.82	0	Fuzzy

Segment ID	N	Mean	Median	SD	Min	Max	Missing	Match category
19	24	42.42	36.36	13.31	27.27	81.82	0	MT
67	24	43.38	44.12	8.81	35.29	64.71	0	Fuzzy
34	24	43.53	43.1	6.33	27.59	58.62	0	MT
23	24	43.75	50	17.54	14.29	64.29	0	Fuzzy
137	24	43.75	43.75	6.65	31.25	56.25	0	Fuzzy
102	24	44.64	42.86	4.83	42.86	57.14	0	Fuzzy
77	24	44.79	50	13.63	18.75	62.5	0	MT
118	24	45.04	47.62	13.32	23.81	71.43	0	MT
95	24	47.22	44.44	16.14	22.22	77.78	0	Fuzzy
129	24	48.51	46.43	9.41	35.71	67.86	0	MT
74	24	50.52	50	5.8	37.5	75	0	Fuzzy
66	24	53.61	53.33	12.81	33.33	86.67	0	Fuzzy
4	24	56.25	50	29.82	0	100	0	Fuzzy
78	24	57.29	59.38	10.85	31.25	68.75	0	MT
135	24	57.69	61.54	16.67	15.39	69.23	0	MT
28	24	61.46	62.5	14.24	0	75	0	MT
36	24	61.81	66.67	25.29	0	100	0	MT

Table 18: TER descriptive data per segment

If the TER score is examined at a segment level (24 versions of each segment) we note that translators did not make similar changes. There are only three segments that were processed in exactly the same way by the 24 translators. These are the segments that have a standard deviation of 0 (segments 65, 86 and 92 from the Fuzzy match category) and that are marked in bold in Table 18, the rest of the segments have different ranges, indicating different TER values, and therefore different edits made by translators.

The following examples illustrate this point:

Segment 65

Source: Click Yes to overwrite the existing analysis.

Fuzzy match: Haga clic en Sí para sobrescribir el **documento** existente.

Edited text: Haga clic en Sí para sobrescribir el **análisis** existente.

Segment 86

Source: Click OK to return to the Share dialog box.

Fuzzy match: Haga clic en Aceptar para volver al cuadro de diálogo **Propiedades**.

Edited text: Haga clic en Aceptar para volver al cuadro de diálogo **Compartir**.

Segment 92

Source: Add objects to the document's dataset and format the second layout.

Fuzzy match: Agregue objetos al conjunto de datos del documento y dé formato al segundo diseño (**¿Cómo?**)

Edited text: Agregue objetos al conjunto de datos del documento y dé formato al segundo diseño.

These three examples show relatively uncomplicated changes to make in order to match the target text to the source text (word substitution or word deletion) and this might explain why the 24 translators made the exact same edit.

The segments translated with MT also show wide ranges in the TER value per segment and there are no segments with the same TER score. This indicates that there were different edits made by all 24 translators in each individual segment. The ones that are more homogeneous, however, are segment 12 (TER range=9.09), segment 27 (TER range=5.26) and 42 (TER range=6.25).

Segment 12

Source: The auto text information will appear inside the text field.

MT: El texto automático información se mostrará dentro del campo de texto.

Post-edited text: La información **de** texto automático **se mostrará** dentro del campo de texto.

La información **del** texto automático **se mostrará** dentro del campo de texto.

La información **del** texto automático **aparecerá** dentro del campo de texto.

La información **de** texto automático **aparecerá** dentro del campo de texto.

In this first instance, the MT is post-edited in four different versions. The word order in the proposed segment was wrong and translators post-edited the text to reflect the source in different forms. In this case, the post-editors present different correct alternatives and also Accuracy errors in two instances.

Segment 27

Source: Add thresholds to an analysis, to change the display of data based on the value of a metric.

Target: Agregue umbrales a un análisis, para cambiar la visualización de los datos en función del valor de un indicador.

Post-edited text: Agregue umbrales a un análisis para cambiar la visualización de los datos en función del valor de un indicador.

Agregue umbrales a un análisis, para cambiar la visualización de los datos en función del valor de un indicador.

In this instance there are two alternatives: one has been post-edited and the other has been left unchanged (leaving an error in the target text (the comma after “análisis”, although this could be consider a minor grammatical error that does not alter the meaning of the source text).

Segment 42

Source: Click the Export icon to the right of the results you want to export.

MT: Haga clic en el icono Exportar situado a la derecha de los resultados que desea exportar.

Post-edited text: Haga clic en el icono Exportar situado a la derecha de los resultados que desea exportar.

Haga clic en el **ícono** Exportar situado a la derecha de los resultados que desea exportar.

Haga clic en el icono Exportar, situado a la derecha de los resultados que desea exportar.

Here the MT is a perfect match. However, one translator has corrected wrongly a term that was in the glossary (“ícono” was changed to “ícono”) and another translator opted to add a comma, which is also a correct variation although it is not compulsory according to Spanish grammar.

At the opposite end, segments 4 (Fuzzy match), 36 (MT match) and 90 (Fuzzy match) presented high deviations and therefore stronger disagreements among translators: some translators did not make any edit, while others made several to the same segment.

Segment 4:

Source Text: The name of the shortcut is updated.

Fuzzy match: El nombre del objeto se actualizará.

Post-edited text: El nombre del acceso directo se actualizará.

El nombre del acceso directo se actualiza.

Se actualiza el nombre del acceso directo.

El nombre del acceso directo se ha actualizado.

El nombre del acceso rápido está actualizado.

El nombre del acceso directo está actualizado.

Se actualizó el nombre del método abreviado.

Se actualizará el nombre del acceso directo.

Segment 36

Source Text: More than one instance is created.

MT: Más de una instancia se crea.

Post-edited text: Más de una instancia se crea

Se crea más de una instancia

Se ha creado más de una instancia

Se creará más de una instancia.

Se crea más de una instancia.

Segment 90

Source Text: These options are listed below as “PDF printing enabled”.

Fuzzy match: Estas opciones aparecerán enumeradas como “impresión DHTML habilitada”.

Edited text: Estas opciones aparecerán enumeradas a continuación como "impresión en PDF habilitada".

Estas opciones aparecerán enumeradas como "impresión en PDF habilitada".

Estas opciones aparecerán abajo enumeradas como "impresión PDF habilitada".

Estas opciones se enumeran a continuación como "impresión en PDF habilitada".

Estas opciones aparecen enumeradas a continuación como "impresión en PDF habilitada".

Estas opciones aparecerán enumeradas como "impresión PDF habilitada".

Estas opciones se detallan a continuación como "impresión de PDF habilitada".

Estas opciones aparecen enumeradas abajo como "impresión DHTML habilitada".

Estas opciones aparecerán enumeradas más abajo como "impresión en PDF habilitada".

Estas opciones aparecerán enumeradas como "impresión DHTML habilitada".

As we can see, in these samples there are several variations of the same proposal. On occasions these variations contain errors (for example, not following the glossary provided or not reflecting the source text exactly) and on other occasions they are correct renderings of the same source text. Once the source text needed to be corrected, translators chose different alternatives, some introduced or left errors, but others simply provided different alternatives.

Although translators made different number of edits, at segment level, was their speed similar (Words per minute)? It is important to look at the relationship between processing speed and number of edits at segment level.

There is a Pearson correlation of -0.42636 ($p = <.0001$) between TER and Words per minute (speed), as shown in Figure 20.

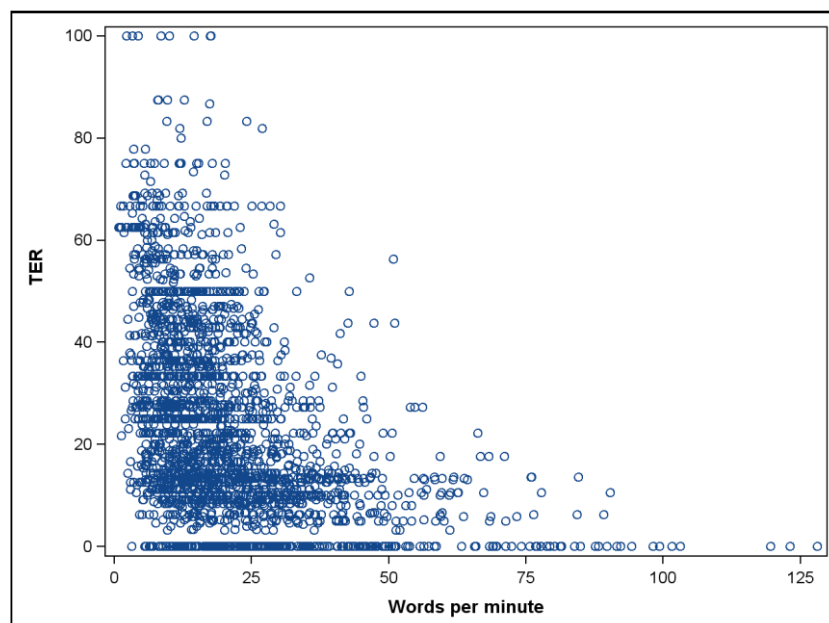


Figure 20: TER and Words per minute

The higher the TER score, the lower the processing speed of that particular segment tends to be, thus indicating that overall the number of edits did affect the speed per segment. However, this relationship is not linear. In order to explore this relation further, a categorization at segment level of the variable *Words per minute* is made as follows:

- Less than 10 words per minute.
- 10 words per minute or more, less than 20 words per minute.
- 20 words per minute or more, less than 30 words per minute.
- 30 words per minute or more.

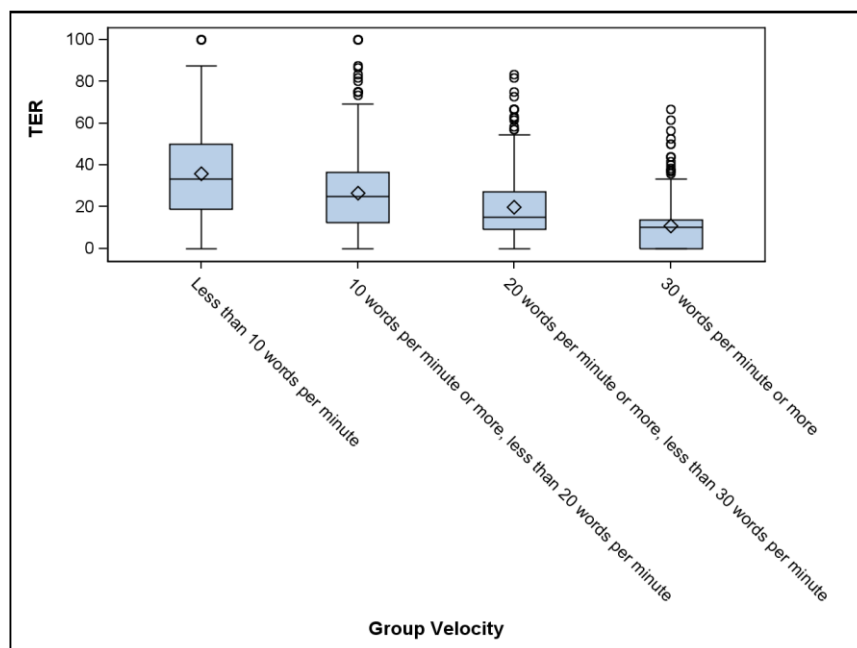


Figure 21: TER vs. Categorized Words per minute

Figure 21 shows that for the different groups the TER scores changes: the score is lower at higher speeds and higher at lower speeds. For the variable *TER* there are statistically significant differences ($F=203.58$ and $p<0.0001$). However, as O'Brien pointed out (2011) some segments that have a higher processing speed (see outliers in 30 words per minute) might still have a high TER score (over 60) and segments that have a lower processing speed (see end of range Less than 10 words per minute) might have a low TER score (0). Figure 22 shows the estimated mean per translator according to these four categories (group velocity).

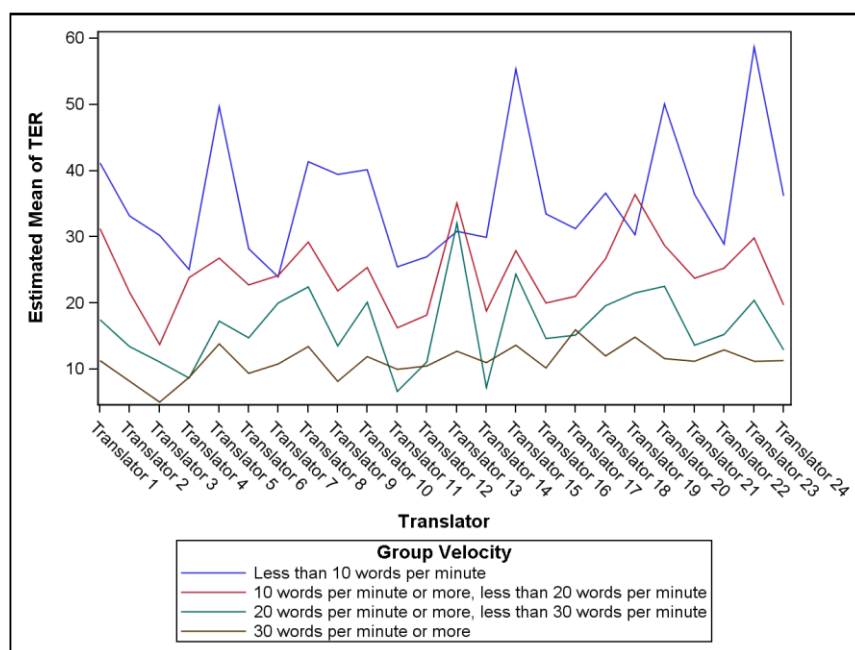


Figure 22: Estimated Mean of TER and Group Velocity

If we turn to the performance of each individual translator, at high speeds the TER values are similar, but there are pronounced variations at medium and slow speeds. It would appear that once a segment needs to be edited, translators behave in different ways, maybe taking the opportunity to introduce other edits that could be preferential, or retranslating the whole segment. At high speeds, once the translation is deemed acceptable, the number of edits is similar among translators. Translator 13 (as seen in Figure 22) has an estimated mean value that is quite similar in all speeds below 30 words per minute. This means that she made a similar number of edits regardless of the speed. Translators 11 and 14 made fewer edits when processing at 20 words per minute than when they did when processing at 30 words per minute. We can conclude that although there is a correlation between high speeds and low TER scores, this is not always the case at segment level or per translator.

These results are in line with Offersgaard's (2008) and O'Brien's (2011) findings on the correlation between automatic metrics and post-editing speed, and the questions posed about the accuracy at a segment level. We further observe in this present study that at higher speeds translators seem to have more homogenous numbers of edits than at medium or lower speeds, and that not all translators behave in the same way with respect to speed and TER scores.

We still have one final question regarding edits: Were faster translators making fewer changes than slower translators? Table 19 shows the translators and their TER

scores distributed according to their speed groups. In the productivity section we saw that translators showed different processing speeds, and we grouped translators according to their speed when translating No match. Speed group 1 is the slowest group when translating No match segments and Speed group 3 is the fastest.

Translator	Speed group 1	Speed group 2	Speed group 3	Total
Translator 01			24.15	24.15
Translator 02	26.22			26.22
Translator 03	24.47			24.47
Translator 04	25.46			25.46
Translator 05			21.12	21.12
Translator 06	22.65			22.65
Translator 07			21.58	21.58
Translator 08		26.79		26.79
Translator 09	28.58			28.58
Translator 10			23.31	23.31
Translator 11	22.35			22.35
Translator 12	23.47			23.47
Translator 13			21.52	21.52
Translator 14	25.69			25.69
Translator 15			24.04	24.04
Translator 16		21.38		21.38
Translator 17		23.82		23.82
Translator 18		23.85		23.85
Translator 19			26.55	26.55
Translator 20			28.58	28.58
Translator 21			22.87	22.87
Translator 22	25.72			25.72
Translator 23			25.80	25.80
Translator 24		21.38		21.38
Average	24.96	23.44	23.95	24.22

Table 19: Global TER according to Speed group

In Table 19, the overall mean TER scores, that is, including Fuzzy and MT matches, are presented according to the Speed group of each translator. We observe that the differences are not pronounced depending on the Speed group even though Speed group 1 has a slight higher average. Figure 23 illustrates better the similarities in TER scores among translators.

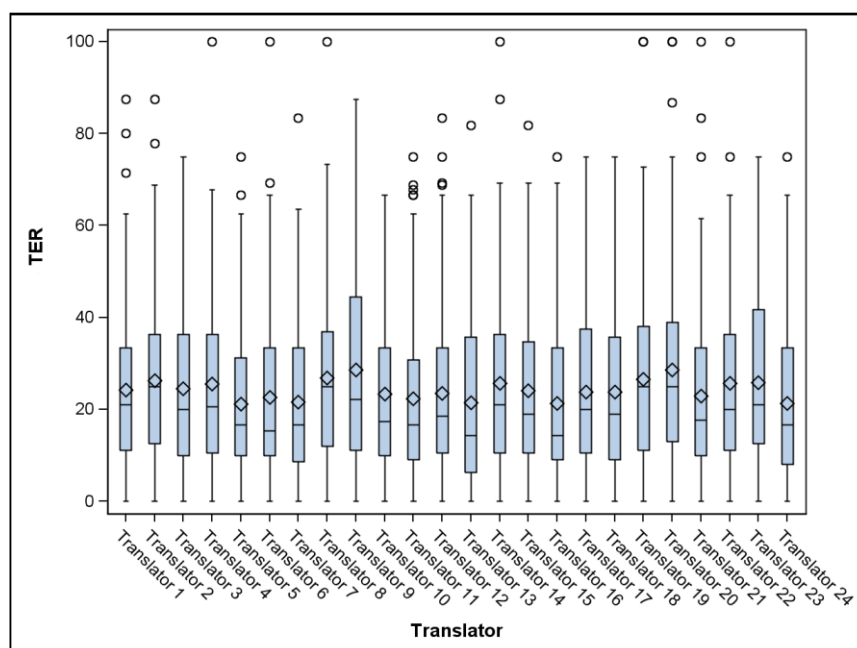


Figure 23: Overall TER values per translator

The TER score does not appear to be related to the speed of the translators as Tatsumi (2009) pointed out in her study, also in Tatsumi and Roturier (2010), and Plitt and Masselot (2010) translators with different processing speeds made similar amount of changes.

In conclusion, the TER scores are statistically significantly different in Fuzzy and MT match categories, and this score is lower in MT matches, meaning that fewer edits were made in these segments than in the Fuzzy matches. If the data are analyzed per translator, we observe that there are statistically significant differences of TER between Fuzzy and MT match in only two translators. Moreover, the TER scores seem to correlate with word processing speed, although this is not true for every segment. Finally, translators show more similarities in TER scores at high speeds than at lower speeds, where more variability in the scores is seen.

4.6. Conclusions on productivity

We have tested our hypothesis and observed that the processing speed for MT and for Fuzzy matches (in the 85-94 percent range) do not present statistically significant differences, although there are statistically significant differences between these two type of matches and the No match category. This seems to suggest that translators gain significant productivity when using translation proposals rather than when translating on their own. The majority of participants in this study show a higher benefit when using MT matches but this value is not significantly higher than when using Fuzzy matches. At the same time, and perhaps this is the reason for the higher values when using MT, we observe that the standard deviation for MT is higher, indicating that some segments require very few edits while others require considerable work. The translators show more homogeneity in speed when translating without any proposal. The data dispersion is nevertheless quite pronounced, with very high standard deviations and great differences between maximum and minimum values, indicating that translators processed the same segments at considerably different speeds. The translators show an increase in productivity when working with Fuzzy and MT matches, but this productivity varies considerably depending on the segments and on the translators. The time savings associated with this productivity also vary considerably (from 6 to 60 percent), and this indicates that some translators benefit more than others from these translation proposals. The average time savings for these translators (32 percent for Fuzzy matches and 37 percent for MT matches) is lower than the average 40 percent assumed for this type of matches (85-94 fuzzy matches) in the industry but this might be explained in some cases by the use of a different tool that does not highlight the changes in the Fuzzy match segments.

When we look at our sub-hypothesis, we observe that translators within different speed groups show different benefits when using MT and Fuzzy match segments: the faster translators (Group 3) seem to benefit less, and the “medium” speed translators more (Group 2). However, no statistically significant differences are observed between the three groups. If we take as reference the speed of each individual translator in the No match category, we observe statistically significant differences when using MT or Fuzzy matches. This indicates that if we take as a starting point the speed at which a translator processes the No match segments (intrinsic speed), and not the absolute

measures (all segments per category), MT segments appear to be processed significantly faster than Fuzzy matches in this project.

Lastly, if we look at edits made in the proposed text using a TER indicator, translators appear to make more edits in Fuzzy match than MT match segments thus indicating that certain MT segments could be perfect matches. This clearly indicates that the initial quality of the MT output significantly defines the productivity that a translator can achieve. If we examine the TER score (that measures the number of specific edits made in all segments), statistically significant differences are found between the Fuzzy and MT matches. However, only two translators out of the 24 show statistically significant differences in the TER scores, if Fuzzy and MT match categories are analyzed. There is a correlation between processing speed and number of edits per segments. The higher the number of edits, the lower the processing speeds of those segments. However, this might not be true for individual segments. Moreover, translators show more consistency in the number of edits when working at higher speeds than when working at medium or lower speeds, where translators appear to make different numbers of edits, thus showing less consistency. We also observe that at segment level translators present more disagreement than agreement on the types of edits to make, despite the fact that some of these are valid renderings of the same source text. Finally, the TER score for translators in different speed groups is not remarkably different, showing that the number of edits is not necessarily related to the processing speed of each translator.

Chapter 5: Quality results

In this chapter we will test our second hypothesis, which states that the final quality of the revised target segments translated using MT technology is higher, if measured in number of errors, than the final quality of revised Fuzzy match segments, and lower than the final quality of revised No match segments. The quality will be measured according to the number of errors in the final texts: the higher the number of errors, the lower the quality. Firstly, the characteristics of the existing corpus will be described to understand better the subsequent results. Secondly, the reviewers' results will be presented, their agreement (or not) according to error classification and the assessment time for all tests. Thirdly, the results obtained from translators in relation to each category will be examined to see if there is one particularly category where there are more errors as well as significant differences between these categories. Finally, we will test the sub-hypothesis that claims that the translators with higher overall processing speeds when using MT or TM technology will have fewer errors than those with lower processing speeds. In other words, processing speed will be correlated with quality. We are interested in finding out whether translators spent more time on their task and this resulted in fewer errors, and consequently less time would be required to revise their translations, or if translators spent less time but they had more errors in the final target text, or simply whether the processing speed had no effect on the number of errors.

Quality will be measured according to the number of errors present in the final target text. We have classified the errors according to the different segment categories (No match, Fuzzy match and MT match), using the LISA QA standard slightly modified for this project (see Appendix E). Three professional translators reviewed all the samples produced by the 24 translators, using the LISA form as a guideline. The focus is on the number and classification of errors, as the aim of this study is not to establish if a particular translator offers good or poor performance but if the number and type of errors is conditioned by the use of a translation aid and therefore if the errors have an impact on the overall productivity of the translation. That is, we seek to investigate whether the time saved using MT (or TM) does not mean additional time in order to fix errors at a later stage in the localization process.

5.1. The corpus and the type of changes required

Before looking at the errors found after the assignment was completed, we would like to describe briefly the type of corrections necessary in order to make the Fuzzy and MT matches correct renderings of the source text. We need to remember that the translation memory used for this project is the same that was used to train the engine. Therefore, the same original quality applies to both types of segments in principle. The internal technical processes are different but the original quality of the material (corpus) used is the same. After training the engine with the existing TM, the MT output could potentially produce a correct translation of the source text, whereas the matches selected from the translation memory (the range 85 to 94 percent in this case) will require a correction, however small. We know, additionally, that the results of the machine translated output after carrying out the human evaluation (see section 3.4.3.4) were rated as having a score of 4.5 out of 5 and this would give an indication of the high quality of the output and indirectly of the translation memory used to train the engine. This does not mean, however, that all the matches in the original translation memory are correctly translated or that all the terminology complies with the glossary provided – it just means that the overall quality is quite acceptable. Some examples of the changes required in both types of segments are presented below in order to understand better the task of the translators, and subsequently of the reviewers.

As was explained above (see section 3.4.3.5), the Fuzzy match category only contains segments from the 85 to the 94 percent match range. This means that the changes required would normally affect a few words that have changed in the source text and thus, needs to be deleted or shifted (see Fuzzy match in Appendix A). However, the TM shows a percentage comparing the old source text and the new source text. If there are changes in the target terminology (Spanish) proposed by the customer, for example, the TM system might show a high level of match (at source-text level) but several changes might be required (at target-text level). Moreover, it could be that the “old” translated segments stored in the memory (and proposed to the translator) contain linguistic errors and those would need to be changed in the target text. Let us look at one example from our corpus per level of match:

85 percent match

Source text:

To display a report as a widget

TM proposed text:

Para visualizar un informe como un widget **de mapa**

In this proposal, the source text does not refer to a “map” while the proposal does, so the translator would have to delete this reference to reflect the change in the source.

88 percent match

Source text:

For example, you add Customer Region, **Profit**, and Revenue to the widget.

TM proposed text:

Por ejemplo, agregue la región del cliente, el **trimestre** y los ingresos al widget.

In this proposal, the source text refers to Profit while the proposal refers to Quarter (*trimestre*). This would need to be changed. Moreover, the glossary shows these software options in upper case and therefore translators will need to change the case and remove the articles. Often translation memories contain “old” solutions that might appear corrected in the updated terminological reference (in this case the glossary).

93 percent match

To add an object displayed in the **MDX** Objects list to the report

Proposed text:

Para agregar un objeto de la lista Objetos **de informe** al informe

In this case, the source text contains the additional term, “MDX”, which would need to be added to the target and others that would need to be deleted, “de informe”.

These changes are just examples but they summarize the types of proposals that translators were presented with and the types of corrections that the reviewers would be expecting to find when examining the final text.

For the MT match category, we will give some examples showing those segments that require, in our view, few edits, and those that require a certain number of corrections to make them equivalent to the source text.

Source text:

This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.

MT proposed text:

En este procedimiento se da por hecho que ya se ha creado un análisis y **ha agregado** al menos dos atributos al panel **de filtros**.

In this example, the MT segment has a problem related to verb agreements, the first part of the sentence uses an impersonal (which is correct) but the second one uses the second person (which is also correct) but one of these two verb forms should be modified to be consistent. Further, the glossary contains a translation for “Filters panel” that does not match the proposal. This should be changed as well.

Source text:

The waiting time is not an integer

MT proposed text:

El tiempo de espera no es un número entero

In this example, the proposal is equivalent to the source text and, a priori, no edits are required.

Source text:

Cube growth check frequency (in mins).

MT proposed text:

Cubo de crecimiento de verificación frecuencia (en mins).

Here, the word order is wrong in the proposal and the translators would need to rearrange most of the segment to reflect the source text. The translation for “mins” should change to “min”, the symbol used in Spanish.

Again, these are just examples, but as we can see for MT segments, some require few or no edits while others need substantial amounts of rearranging or correcting. The difference in the required post-edit change could mean different results in the final quality of the text. It is important to reiterate at this point that the post-editors did not know the origin of the segments (MT or Fuzzy) and obviously whether these segments were full (100 percent) or fuzzy matches (54-99 percent). They were just presented with a proposal and they were instructed to edit it according to the instructions given (see Appendix C).

5.2. The review

As we explained in section 3.5, the reviewers sent back 24 LISA forms, 24 edited Word documents (with tracked changes) and one timesheet with the registered time employed to correct each text and to complete each form. With this information in hand, all errors were transferred to three Excel documents. See section II.1 Statistical analysis for a full description of the databases and the statistical analysis for the quality chapter.

5.2.1. Revision time

Before looking at the actual quality results from the translators, it is necessary to establish if there was agreement in terms of corrections and time among the three reviewers. Figure 24 presents the time reviewers took to complete the task. Each reviewer had 24 translations and that meant a total of 50,976 words to review. They were dealing, however, with one text repeated 24 times. When queried, the reviewers confirmed that they had reviewed in strict order from 1 to 24, although on occasions they went back to change some corrections in previous translations.

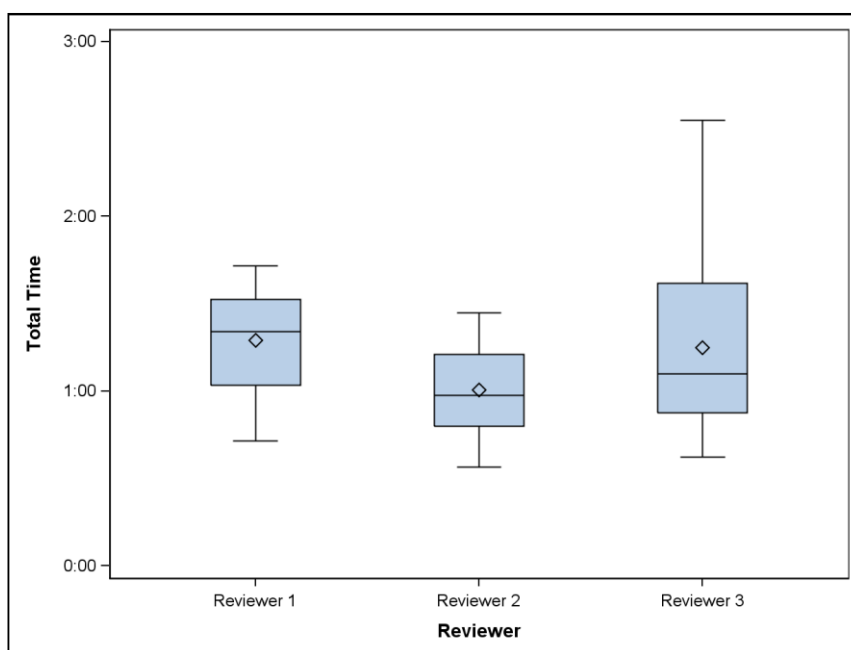


Figure 24: Total reviewing time

The times are different for the three reviewers. Reviewer 1 took 30 hours and 58 minutes to review all 24 tests; Reviewer 2 took 24 hours and 10 minutes and Reviewer 3 took 29 hours 59 minutes and 18 seconds. As explained in section 3.5, the time invested by the reviewers in clarifying certain errors to us when we were transferring

results is not included in these figures, but only the time the three reviewers took to go through the bulk of the translations, marked the Word documents and transfer the results to the LISA QA forms. The times of Reviewers 1 and 3 are almost equal; their approaches to the tasks, however, seem to be different, since the mean and median values are different, and Reviewer 3 has a wider range of time. This reviewer took longer initially than the other two and then she gained speed as the text became familiar. In the end, however, she corrected certain texts quicker than Reviewer 1 (this can be appreciated in the minimum value for Reviewer 3, which is below 1). Although Reviewer 2 was faster, she forgot to include the types of error in the Word files and she had to go back and correct the mistake, adding the time to the Excel form, as explained in section 3.5. It could be possible that the time keeping was somewhat distorted as a result, but it could also be that she was simply faster at correcting all the texts. Let us examine these results descriptively in Table 20 to gather more information.

Reviewer	N	Mean	Median	SD	Min	Max
Reviewer 1	24	1:17:25	1:20:30	0:16:38	0:43:00	1:43:00
Reviewer 2	24	1:00:25	0:58:30	0:15:25	0:34:00	1:27:00
Reviewer 3	24	1:14:58	1:06:00	0:29:29	0:37:20	2:33:00

Reviewer	N	Min	Q 1	Median	Q 3	Max	Range	Q Range
Reviewer 1	24	0:43:00	1:02:00	1:20:30	1:31:30	1:43:00	1:00:00	0:29:30
Reviewer 2	24	0:34:00	0:48:00	0:58:30	1:12:30	1:27:00	0:53:00	0:24:30
Reviewer 3	24	0:37:20	0:52:30	1:06:00	1:37:00	2:33:00	1:55:40	0:44:30

Table 20: Descriptive analysis of Review time

Reviewer 2 was the fastest reviewer if the mean value is considered, but also if the minimum and maximum values are considered. This means that her longest review took 1 hour and 27 minutes (for Translator 3) and the shortest was 34 minutes (for Translator 23). Reviewer 1 has a higher mean value than Reviewer 3 but he was more consistent regarding time over the 24 translations, since the minimum value is 43 minutes (for Translator 20) and the maximum 1 hour and 43 minutes (for Translator 3). Reviewer 3 has a wider range (44:30) because she took a maximum of 2 hours and 33 minutes (for Translator 1, the first test) and a minimum of 37 minutes and 20 seconds (for Translator 20). Therefore, Reviewer 2 appears to be faster overall and also per individual test. Reviewer 1 is the slowest overall but more consistent with each individual text in terms of timing. Figure 25 and Table 21 show the number of words corrected per minute.

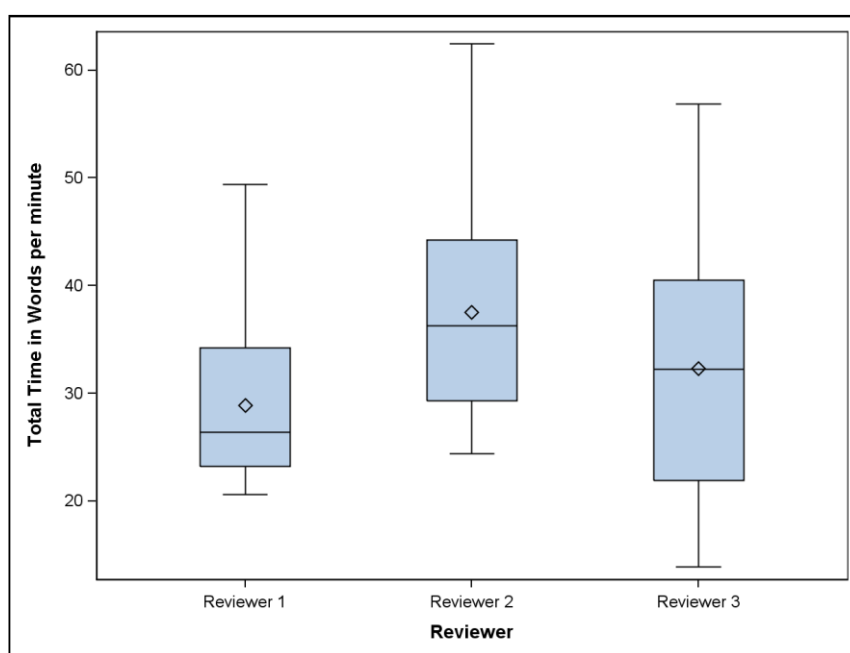


Figure 25: Total reviewing time in words per minute

Reviewer	N	Mean	Median	SD	Min	Max
Reviewer 1	24	28.91	26.39	7.36	20.62	49.40
Reviewer 2	24	37.50	36.31	9.94	24.41	62.47
Reviewer 3	24	32.27	32.25	11.24	13.88	56.89

Reviewer	N	Min	Quartile 1	Median	Quartile 3	Max	Range	Quartile Range
Reviewer 1	24	20.62	23.21	26.39	34.26	49.40	28.77	11.04
Reviewer 2	24	24.41	29.31	36.31	44.25	62.47	38.06	14.94
Reviewer 3	24	13.88	21.90	32.25	40.49	56.89	43.01	18.59

Table 21: Descriptive analysis of Review time in WPM

We observe, as above, that Reviewer 2 has the fastest reviewing time, followed by 3 and then by 1. Reviewer 1 seems to review at a more similar pace throughout the test (the standard deviation is 7.36) and Reviewer 3, with a deviation of 11.24, has a wider range of 18.59. Again, the minimum value for Reviewer 3 is quite low at 13.88 words reviewed per minute but the maximum is the second highest at 56.89 words reviewed per minute. If the data is modeled with repeated measures taking *logarithm of Total time in words per minute* as the response variable and *Reviewer* as the explanatory variable, there are statistically significant differences among the three reviewers ($F=17.30$; $p<0.0001$). However, there are no statistically significant differences between Reviewers 1 and 3 with regards to time.

If the results obtained during the assignment are extrapolated to 8 hours (the average working hours considered when allocating a reviewing task), we have the mean values shown in Table 22.

Reviewer	Total words reviewed	Total review time	Mean reviewed in 8 h
Reviewer 1	50,976	30:58:00	13,169
Reviewer 2	50,976	24:10:00	16,875
Reviewer 3	50,976	29:59:18	13,599

Table 22: Mean of total words reviewed in 8 hours

The number of words is greater than the averages used in the localization industry (between 5000 and 10000 words) but we also have to take into consideration that, in this case, although the total volume of words is 50,976, the reviewers were looking at the same text of 2,100 words repeated 24 times. However, in this case, additional tasks were requested: to track changes in Word, to track time in timesheets, to transfer errors per Match category, and this could also add to the overall time spent reviewing.

We shall now return to the number of errors to examine if there was agreement among the three reviewers.

5.2.2. Number of errors in review

First of all, let us examine those segments that were highlighted by reviewers as containing an error. For this, an error indicator is defined:

- 0 means that there are no errors
- 1 means that there are errors

To examine the degree of agreement among reviewers, another variable is defined with the following values:

- There are no differences among translators
- Reviewer 1 does not agree with Reviewers 2 and 3
- Reviewer 2 does not agree with Reviewers 1 and 3
- Reviewer 3 does not agree with Reviewers 1 and 2.

The error indicator will only highlight whether those segments contain an error or not, not the number of errors per segment. However, it is important to note that, out of the 10,728 segments, the reviewers marked mostly 1 error per segment; only in two segments did they mark 3 errors, and 12 segments out of 10,728 they marked 2 errors

(10,728 segments if we consider that there were 149 segments times 24 translators times three reviewers). This should give us an idea of the degree of agreement on the number of segments in which corrections had to be made.

Reviewers' agreement	Fuzzy match		MT match		No match		All	
	N	%	N	%	N	%	N	%
No differences	953	79.42	866	73.64	794	66.17	2613	73.07
Reviewer 1 does not agree	64	5.33	109	9.27	142	11.83	315	8.81
Reviewer 2 does not agree	103	8.58	103	8.76	137	11.42	343	9.59
Reviewer 3 does not agree	80	6.67	98	8.33	127	10.58	305	8.53

Table 23: Percentage of error indicator

For the 24 translators there are a total of 3,576 segments (149 segments times 24 translators). The reviewers agree on 73.07 percent of all segments (2,613) and they disagree on 26.93 percent (963 segments). There is more agreement on the Fuzzy matches (79.42 percent) and less on the No matches (66.17 percent). Since we transferred the results from the forms and Word texts from the reviewers, we are aware that the individual corrections are not exactly the same. The data above indicate solely in which segments there was an error marked; it does not tell us if there was agreement on the number of errors or in the type of errors marked.

Reviewer	Fuzzy match		MT match		No match		All	
	N	Error #	N	Error #	N	Error #	N	Error #
Reviewer 1	1200	187	1176	171	1200	309	3576	667
Reviewer 2	1200	206	1176	173	1200	287	3576	666
Reviewer 3	1200	149	1176	232	1200	352	3576	733

Table 24: Total number of errors

Table 24 shows the absolute numbers of segments containing errors. The number of words is not considered and, since it was slightly different per category, the ratio of errors per word might differ. Reviewers 1 and 2 differ in total number of segments containing errors by only one, yet these are distributed differently across two categories (No match and Fuzzy match in particular) and they show very similar results in the MT match category, a difference of only two errors. Reviewer 3 shows a higher number of errors than the other two reviewers in all categories but Fuzzy matches, although they all agree that the No match category has more errors. The number of segments containing errors per translator and category is also different between the three reviewers. See Appendix G for detailed information per translator.

Reviewer	Time/errors	N	Mean	Median	SD	Min	Max
Reviewer 1	Total Time in WPM	24	28.91	26.39	7.36	20.62	49.40
	Total Errors	24	27.79	23.50	12.89	13.00	60.00
Reviewer 2	Total Time in WPM	24	37.50	36.31	9.94	24.41	62.47
	Total Errors	24	27.75	24.00	11.97	14.00	57.00
Reviewer 3	Total Time in WPM	24	32.27	32.25	11.24	13.88	56.89
	Total Errors	24	30.71	23.50	15.19	12.00	64.00

Reviewer	Time/errors	N	Min	Q 1	Median	Q 3	Max	Range	Q Range
Reviewer 1	Time in WPM	24	20.62	23.21	26.39	34.26	49.40	28.77	11.04
	Errors	24	13.00	18.00	23.50	32.00	60.00	47.00	14.00
Reviewer 2	Time in WPM	24	24.41	29.31	36.31	44.25	62.47	38.06	14.94
	Errors	24	14.00	19.00	24.00	35.00	57.00	43.00	16.00
Reviewer 3	Time in WPM	24	13.88	21.90	32.25	40.49	56.89	43.01	18.59
	Errors	24	12.00	19.50	23.50	38.50	64.00	52.00	19.00

Table 25: Descriptive data on errors per reviewer

By simply looking at these descriptive data in Table 25 from the reviewers it can be seen that the results are different. The minimum and maximum values are quite similar. In other words, the reviewers agree on the very poor and very good results. The mean value for Reviewer 3 is slightly higher: she made more corrections overall. For Quartile 1 there is relative agreement, suggesting that it might be easier to agree on the translators that made minimum and median errors (Minimum, Quartile 1 and Median values), than those that made more errors (Quartile 3 and Maximum values) possibly because once there are more errors, one reviewer might decide to correct or adapt more things to his or her taste (as we also saw for the translators when post-editing the text).

To see if there are statistically significant differences, we modeled the data with repeated measures taking the *logarithm of Words per minute* (in this case, words per minute to review) as the response variable and *Total errors* and *Reviewer* as explanatory variables. There are statistically significant differences between the Reviewers ($F=18.00$; $p<0.0001$) and for the *Total errors* ($F=14.39$; $p=0.0004$).

Figure 26 and Figure 27 illustrate these findings, the first figure highlights the time lines, and the second the error lines for the three reviewers.

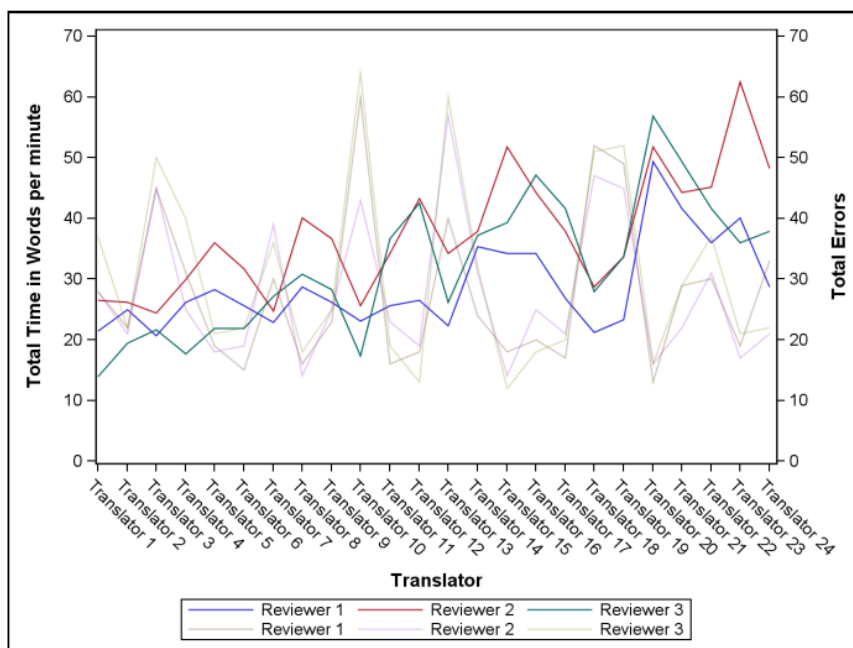


Figure 26: Total time, in words per minute, taken by Reviewers

Figure 26 shows how reviewers gained speed as the 24 texts were reviewed. However, in particular cases, the speed decreases for those translators that have more errors, even at the latest stage of the review (Translators 13, 18, 19, 22 and 24).

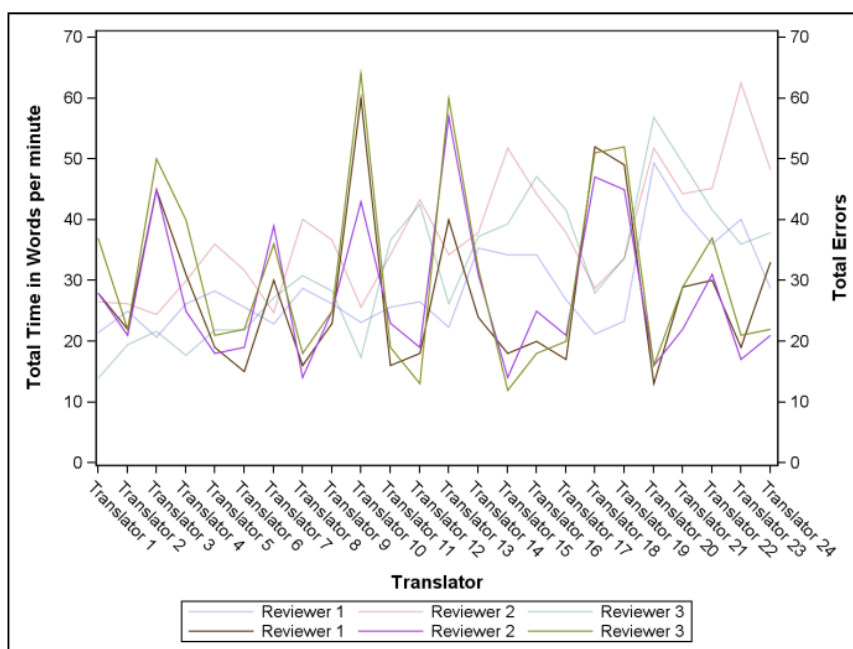


Figure 27: Total Errors Reviewers

In Figure 27, the speed has not affected the number of errors found. There are more errors for Translators 3, 7, 10, 13, 18, 19 and 22 (peaks) and fewer for Translators

2, 5, 6, 8, 12, 15, 16, 17 and 20. The three lines (brown, pink and green) representing the errors marked by the reviewers are not exactly the same per translator.

We know that there is no agreement between the three reviewers, but we have seen that Reviewer 1 and Reviewer 2 had a similar number of global errors. It is important to remember that the data analyzed has a hierarchical structure (that is, there are 24 translators, 149 segments per translator, three Match categories, 8 Types of errors, and three reviewers), the fact that the global number of errors (between Reviewers 1 and 2, for example) is similar does not necessarily mean that the reviewers agree on the errors marked for each translator in each segment for each category. To explore the relationship between reviewers further, we decided to compare their association.

5.2.3. Comparing reviewers

In order to see the agreement or association between the three reviewers in terms of numbers of errors, we classified errors into three categories: Few errors, Average errors, and Many errors. Since the number of errors that the reviewers marked in the three translation categories (Fuzzy, MT and No match) was very different, as we have seen above, we used the first and third quartiles to determine where to divide the data. The results were:

For Fuzzy and MT match:

- Between 0 and 5 errors;
- Between 6 and 10 errors;
- More than 10 errors.

For No match:

- Between 0 and 8 errors;
- Between 9 and 16 errors;
- More than 16 errors.

We calculated the contingency tables and we applied the Kappa coefficient (see Kappa coefficient in Appendix A) per Match category and Reviewer and we observed that for the No match category there is less variability, the Kappa coefficient for the categorization of number of errors is from 0.61 to 0.75 ($p < .0001$), and this means that there is agreement between the three reviewers. On the other hand, for Fuzzy and MT

match, the Kappa measure is more variable and it is between values that show either no agreement ($-0.02 \approx 0$) or a weak agreement (0.43). The only cases where the relation (disagreement in this case) is clear are:

- For Fuzzy match, Reviewer 2 vs. Reviewer 3: There is no agreement between these two reviewers.
- For MT match, Reviewer 2 vs. Reviewer 3: There is no agreement between these two reviewers.

Table 26 illustrates the agreements or associations between the three reviewers according to the Match categories giving the Kappa values.

Match	Contingency table	Kappa	95% Lower	95% Upper	Two-sided Pr > Z
Fuzzy	Table Reviewer_1 * Reviewer_2	0.43	0.14	0.72	0.00
Fuzzy	Table Reviewer_1 * Reviewer_3	0.30	0.02	0.59	0.02
Fuzzy	Table Reviewer_2 * Reviewer_3	-0.02	-0.24	0.20	0.84
MT	Table Reviewer_1 * Reviewer_2	0.36	0.05	0.68	0.01
MT	Table Reviewer_1 * Reviewer_3	0.39	0.12	0.66	0.00
MT	Table Reviewer_2 * Reviewer_3	0.17	-0.12	0.46	0.21
No match	Table Reviewer_1 * Reviewer_2	0.62	0.37	0.87	<.0001
No match	Table Reviewer_1 * Reviewer_3	0.61	0.34	0.88	<.0001
No match	Table Reviewer_2 * Reviewer_3	0.75	0.53	0.97	<.0001

Table 26: Kappa statistical values according to Match category and Reviewer

Therefore, the three reviewers agree on the No match category and Reviewer 1 slightly agrees with Reviewers 2 and 3 in Fuzzy and MT match but Reviewers 2 and 3 do not agree. Perhaps the fact that the reviewers knew the origin of the segments influenced them when marking errors. It is difficult to say. In this analysis, we are comparing them in terms of errors per translator, but still they might disagree per segment or per classification of errors. Below we present some samples taken from the actual reviewed files to illustrate this divergence in the reviewing criteria. The target translations are taken from Translator 15 and the corrections made by each reviewer in the first three segments are shown.

Reviewer 1

Segment-ID	Source-Segment	MT-Target-Segment	Post-edited-target	TM-Match	PE-difference(%)	Type-of-error
4	The name of the shortcut is updated.	El nombre del objeto se actualizará.	El nombre del acceso directo se actualizará.	Fuzzy-match	82,93	
5	You can filter the data displayed in the analysis by selecting options from the Filters panel.		Para de filtrar los datos presentados por en el análisis, se seleccionando las opciones del panel Filtros.	No-match	0	Language(1)
6	This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.	En este procedimiento se da por hecho que ya se ha creado un análisis y ha agregado al menos dos atributos al panel de filtros.	En este procedimiento se da por hecho que ya se ha creado un análisis y que se han agregado al menos dos atributos al panel Filtros.	MT-match	95,02	

Figure 28: Sample 1 correction Reviewer 1

In this case Reviewer 1 only corrected segment 5 as a Language error. The problem appears to be that Translator 15 used a *gerundio* (Spanish gerund) to construct the sentence (No match) and the reviewer found that this was a mistake. Although the *gerundio modal* (used in this sample) is accepted in Spanish (see Real Academia Española 2009 and Fundeu 2012), the *gerundio de consecuencia* in English is wrongly translated in Spanish as a *gerundio*. Therefore, this might be the reason why Reviewer 1 corrected it.

Reviewer 2

Segment-ID	Source-Segment	MT-Target-Segment	Post-edited-target	TM-Match	PE-difference(%)	Type-of-error
4	The name of the shortcut is updated.	El nombre del objeto se actualizará.	El nombre del acceso directo se actualizará.	Fuzzy-match	82,93	
5	You can filter the data displayed in the analysis by selecting options from the Filters panel.		Puede filtrar los datos presentados por el análisis seleccionando las opciones del panel Filtros.	No-match	0	
6	This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.	En este procedimiento se da por hecho que ya se ha creado un análisis y ha agregado al menos dos atributos al panel de filtros.	En este procedimiento se da por hecho que ya se ha creado un análisis y que se han agregado al menos dos atributos al panel Filtros.	MT-match	95,02	

Figure 29: Sample 1 correction Reviewer 2

Reviewer 2 on the other hand did not correct any of the first 3 segments. She deemed the translations correct.

Reviewer 3

SegmentID	Source-Segment	MT-Target-Segment	Post-edited-target	TM-Match	PE-difference(%)	Type-of-error
4	The name of the shortcut is updated.	El nombre del objeto se actualizará.	El nombre del acceso directo se actualizará.	Fuzzy-match	82,93	Accuracy
5	You can filter the data displayed in the analysis by selecting options from the Filters panel.		Para que filtrar los datos presentados en por el análisis, selecciona and las opciones del panel Filtros.	No-match	0	Language
6	This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.	En este procedimiento se da por hecho que ya se ha creado un análisis y ha agregado al menos dos atributos al panel de filtros.	En este procedimiento se da por hecho que ya se ha creado un análisis y que se han agregado al menos dos atributos al panel Filtros.	MT-match	95,02	

Figure 30: Sample 1 correction Reviewer 3

Reviewer 3 highlighted the same error in segment 5 as Reviewer 1 did and also another error in segment 4. The phrase “is updated” was translated as “*se actualizará*” (future) instead of “*se actualiza*” (present), and this was deemed to be an Accuracy error (Fuzzy match). In English, the difference would be between “is updated” and “will be updated”. Reviewers 1 and 2 did not correct this, possibly because the present tense here does not necessarily indicate that the action has finished, and therefore both translations are potentially correct. All reviewers agreed that segment 6 did not require any change: Reviewers 1 and 3 agreed that segment 5 required the same change; Reviewers 1 and 2 agreed that segment 4 did not require any change, but Reviewer 3 disagreed.

Let us look at another sample from Translator 10:

Reviewer 1

Segment ID	Source Segment	MT-Target Segment	Post-edited target	TM-Match	PE-difference(%)	Type of error
4	The name of the shortcut is updated.	El nombre del objeto se actualizará.	El nombre del acceso-rápido-directo está actualizado.	Fuzzy-match	69,05	Terminology(1)
5	You can filter the data displayed in the analysis by selecting options from the Filters panel.		Para poder filtrar los datos que se muestran en el análisis selección de opciones del panel de Filtros.	No match	0	Language(2)
6	This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.	En este procedimiento se da por hecho que ya se ha creado un análisis y ha agregado al menos dos atributos al panel de filtros.	En este procedimiento se da por hecho que ya se ha creado un análisis y se han agregado al menos dos atributos al panel de Filtros.	MT-match	97,69	

Figure 31: Sample 2 correction Reviewer 1

Reviewer 1 has marked a terminology error in segment 4 to make the translation consistent with the glossary. He has also marked two language errors in segment 5. One is related to the present participle we presented earlier, and the other one to a spelling mistake in “*Fritlos*”, which should be spelled correctly as “*Filtros*”. No changes were made in segment 6.

Reviewer 2

Segment ID	Source Segment	MT-Target Segment	Post-edited target	TM-Match	PE-difference(%)	Type of error
4	The name of the shortcut is updated.	El nombre del objeto se actualizará.	El nombre del acceso-rápido-está actualizado se actualiza.	Fuzzy-match	69,05	Overcorrection
5	You can filter the data displayed in the analysis by selecting options from the Filters panel.		Puede filtrar los datos que se muestran en el análisis seleccionando opciones del panel de Filtros.	No match	0	Language
6	This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.	En este procedimiento se da por hecho que ya se ha creado un análisis y ha agregado al menos dos atributos al panel de filtros.	En este procedimiento se da por hecho que ya se ha creado un análisis y se han agregado al menos dos atributos al panel de Filtros.	MT-match	97,69	Language

Figure 32: Sample 2 correction Reviewer 2

Reviewer 2 has corrected two Language errors in segments 5 and 6. In segment 5, we see the same spelling mistake, “*Fritlos*”. In segment 6, the preposition “*de*” is corrected. Reviewer 2 is considering here that “*Filtros*” is a proper noun (the name of the panel) and therefore the noun is in upper case and no preposition is required. This

could also be considered a terminology error since the glossary clearly indicated “*panel Filtros*”. Reviewer 2 has not marked any errors in segment 4 but has identified an overcorrection (see Overcorrection in Appendix A), and in doing so, she has actually inserted a change between “*actualizará*” and “*actualiza*”, perhaps because she thought both were correct options (unlike Reviewer 3 with Translator 15 earlier on).

Reviewer 3

Segment-ID	Source-Segment	MT-Target-Segment	Post-edited target	TM-Match	PE-difference(%)	Type-of-error
4	The name of the shortcut is updated.	El nombre del objeto se actualizará.	El nombre del acceso <u>directo</u> rápido está actualizado.	Fuzzy match	69,05	Terminology
5	You can filter the data displayed in the analysis by selecting options from the Filters panel.		Para <u>de</u> filtrar los datos que se muestran en el análisis, seleccione <u>las</u> <u>de</u> opciones del panel de Filtros.	No match	0	Language
6	This procedure assumes that you have already created an analysis and added at least two attributes to the Filters panel.	En este procedimiento se da por hecho que ya se ha creado un análisis y ha agregado al menos dos atributos al panel de filtros.	En este procedimiento se da por hecho que ya se ha creado un análisis y se han agregado al menos dos atributos al panel de Filtros.	MT match	97,69	

Figure 33: Sample 2 correction Reviewer 3

Reviewer 3 has marked one Terminology error in segment 4 and Language errors in segment 5, like Reviewer 1, except that she inserted a comma (she counted two errors in the LISA form). Reviewers 1 and 3 seem to agree overall on the changes except for a comma, but Reviewer 2 agrees partially on segment 5 and did not agree on segments 4 and 6.

There are many more examples of these agreements and disagreements. We have picked these ones because they represent the type of corrections and disagreements in all 24 texts. On the one hand, it is understandable that by having three reviewers there is an exposure to three different versions, even though great emphasis was placed on not introducing any preferential changes (see Appendix C). On the other hand, we were quite surprised at the disagreements and at some of the changes made. That said, other studies, such as Carl et al. (2011) and García (2010, 2011), have also found disagreement among reviewers. Although not strictly referring to professional reviewers of post-edited segments, but to MT evaluation metrics, Koehn (2012b) also reports on disagreements between MT human evaluators in several studies. The fact that

preferential changes are made and errors might be introduced or left uncorrected during revision has been studied by Brunette (2005), Künzli (2006, 2007), and Mossop (2007b) among others. The aim of our study was not to look at the value added by revision or the different approaches to the revision task, nor to rate reviewers, but to evaluate the final quality according to the number of errors established by independent reviewers. Nevertheless, we do believe that further analysis of these results and reviews could be of interest in a future project.

5.2.4. Global error database vs. Segment error database

We would like to discuss briefly the differences between the Global error database and the Segment error database, and the reason for choosing one or the other depending on the objective of the analysis.

The Global error database contains the errors transferred from all LISA forms. Here, the reviewers were instructed not to count the exact same error twice. For example, if a translator made the mistake of capitalizing all nouns in a software option, not following the glossary, the error was counted once, even though there might be more errors in another segment of a similar nature. Although the instructions might not be clear in most QA models (O'Brien 2012), in our experience in the industry, repeated errors are not counted twice. This is understandable since if there is one translation, for example, with one software option translated wrongly in the whole text (let us say that the glossary was overlooked that one time) and this is repeated 50 times in the file, a global change will suffice to correct this oversight, but if there is another translation with 50 software options translated wrongly (because the glossary was not consulted), then it would take a considerable amount of time to correct 50 different instances in the file. Therefore, it is quite frequent that repeated errors are counted only once. We have used this database to count the errors per translator as well as the relation between errors and speed.

The Segment error database contains all errors, transferred from the Word documents in all segments, including the repeated errors. This database will contain more errors than the Global one, since all the repeated errors are included. We use this database to calculate the percentage of errors in the translations at segment level.

There is an issue that it is important to examine before moving on to the results. Because the reviewers would count one error once (even if repeated) it could be that one of the categories (No match, Fuzzy match or MT match) was more affected by this

methodology than the other two. If an error was spotted in one type of segment (for example No match) that happened to be the first one the reviewer corrected but the exact same error also appeared later on in another type of segment (Fuzzy match), the error would be assigned to the first category (No match). Since reviewers were only supposed to count the errors once in the LISA QA model, they would assign it to the first Match category in the file. Therefore it was important to look at the correlation between these two databases according to the Match category. Figure 34 shows their comparison.

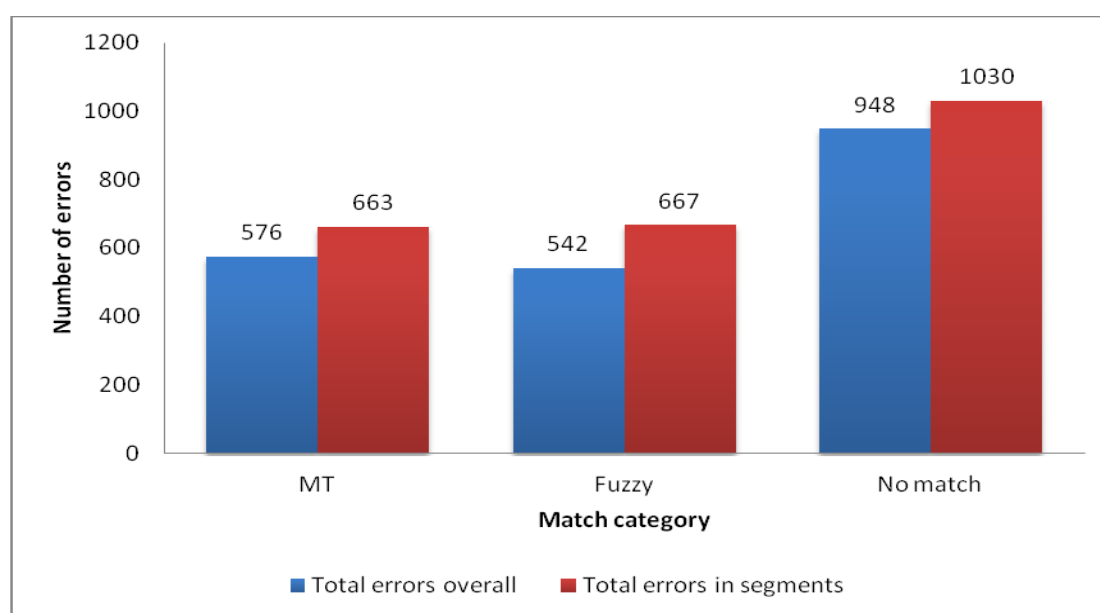


Figure 34: Global error database vs. Segment error database

There are slightly more errors in the Segment error database in the three categories, as can be seen in Table 27. It can also be seen that there are more errors in the No match categories in both databases. There are slightly more errors in the Fuzzy match segments than in the MT match segments in the Segment error database, and there are slightly more errors in the MT match segments than in the Fuzzy match segments in the Global error database. However, these differences are only marginal.

Match	Rev 1		Rev 2		Rev 3	
	Global DB	Segment DB	Global DB	Segment DB	Global DB	Segment DB
Fuzzy match	187	229	206	257	149	181
MT match	171	190	173	212	232	261
No match	309	335	287	329	352	366

Table 27: Global error database vs. Segment error database

Appendix G shows a full table per Translator, Match category and Reviewer. Most of the differences are in Terminology and Language errors. This is logical, since these are the types of errors that would tend to be repetitive errors, normally affecting one word, as opposed to a Mistranslation error, which would concern one specific sentence, normally affecting several words or an expression where it is unlikely that this same pattern would be repeated in another segment.

Differences in number of errors		N	Mean	Median	SD	Min	Max
Differences	Accuracy	216	-0.01	0.00	0.10	-1.00	0.00
Differences	Consistency	216	0.00	0.00	0.00	0.00	0.00
Differences	Country	216	0.00	0.00	0.00	0.00	0.00
Differences	Format	216	-0.00	0.00	0.07	-1.00	0.00
Differences	Language	216	-0.35	0.00	0.82	-6.00	0.00
Differences	Mistranslation	216	-0.01	0.00	0.12	-1.00	0.00
Differences	Style	216	-0.06	0.00	0.29	-2.00	0.00
Differences	Terminology	216	-0.90	0.00	2.02	-12.00	0.00
Differences Total		216	-1.34	-1.00	2.26	-14.00	0.00

Table 28: Differences between databases according to error typology

Table 28 shows that of the 216 items (24 translators times 3 reviewers times 3 categories) in the two tables all the differences are negative or 0. The 0 means that there are no differences in number of errors between the two, and the negative value means that there are more errors in the Segment error database for one particular category and translator. The highest differences are in Terminology, mainly, and Language. These results show that using one or the other database to calculate the number of errors per category in the target texts will not affect the final conclusion on number of errors per category.

5.3. Errors at segment level

In order to see the number of segments that contain errors and those that did not, the data from the Segment database were used and the following values obtained (Figure 35).

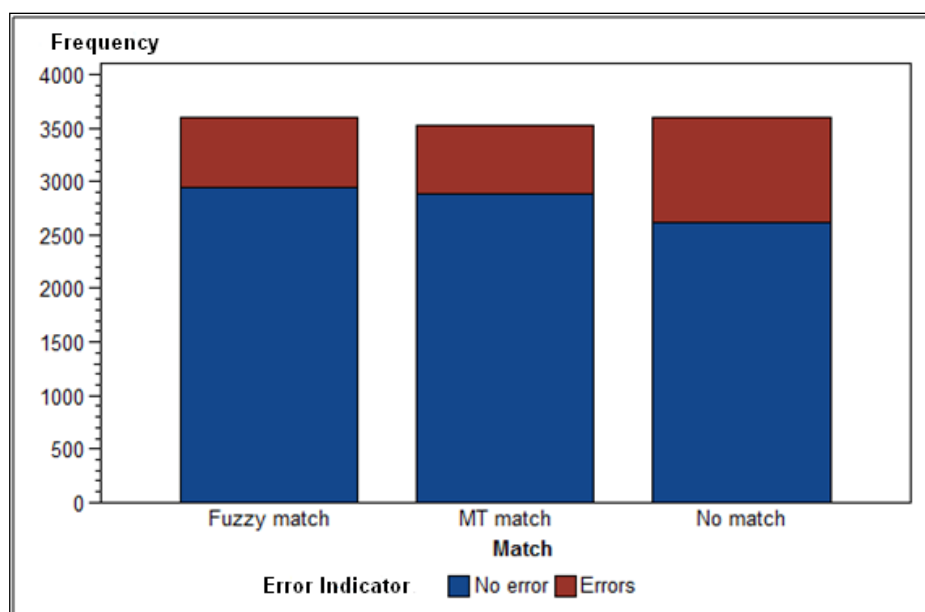


Figure 35: Error indicator per category

Match	No errors		Errors	
	N	%	N	%
Fuzzy match	2946	81.83	654	18.17
MT match	2889	81.89	639	18.11
No match	2618	72.72	982	27.28
All	8453	78.79	2275	21.21

Table 29: Error indicator per category

Almost 80 percent of all segments reviewed did not contain any errors (78.79 percent) and 21 needed some kind of correction. This figure is quite revealing if we consider that when we calculated the TER for the translators, 89.65 percent of all segments were edited, so this means that translators edited almost 90 percent of all segments, and that reviewers found the majority of these edits were appropriate, since they only corrected 21 percent of all segments. The other interesting figure is that the Fuzzy and MT matches have almost identical percentages of segments that were corrected by reviewers, 18.17 and 18.11 percent respectively. Even the No match category, which contains more segments edited (27.28 percent), still has 72.72 percent of segments that did not require any change. Of course, the quality of a translation is not measured only in terms of those segments that are not corrected but according to the type and number of errors as well. Nevertheless, judging from these results, translators delivered quite a high percentage of error-free segments according to the reviewers. Attention should be drawn to the fact that although this task was reviewed by three

professional translators with experience, this does not mean that the reviewers cannot make mistakes or insert preferential changes.

We modeled the data with repeated measures, taking *Error indicator* as the response variable, to observe the possibility of making an error in a segment depending on the category. Statistically significant differences are observed in the proportion of errors in the three different types of translations Fuzzy match, MT match and No match ($F=62.15$, $p<0.0001$). From this model, the estimated odds of “making a mistake” are calculated. Table 30 shows this value with the corresponding confidence intervals of 95 percent.

Match	Odds error	Lower	Upper
Fuzzy match	0.20	0.24	0.17
MT match	0.20	0.24	0.17
No match	0.35	0.41	0.30

Table 30: Odds of “making a mistake”

In Fuzzy match, the estimated odds of “making a mistake” are 0.2, that is, the probability of making a mistake is 0.2 times the probability of not making a mistake. In MT match, the estimated odds are also 0.2 times. In No match, the estimated odds are 0.35 times the probability of not making a mistake. This Match category is the one with the highest probability of making an error. The Odds *Ratio* tells us the relation between No match and Fuzzy match and No match and MT match and its confidence levels, as shown in Table 31.

Match	Match	Odds Ratio	Lower	Upper
No match	Fuzzy match	1.75	1.56	1.96
No match	MT match	1.75	1.56	1.97

Table 31: Estimated Odds ratio per category

The Odds Ratio estimation of No match with regards Fuzzy match is 1.75, that is, the odds of “making a mistake” in No match is 1.75 times the odds of “making a mistake” in Fuzzy match. The Odds Ratio of No match with regards the MT match is 1.75. In other words, the odds of “making a mistake” in No match is 1.75 times the odds of making a mistake in the MT match category.

Table 32 shows the error indicator according to each reviewer instead of per Match category as we saw above.

Reviewer	No error		Errors		Total segments
	N	%	N	%	
Reviewer 1	2855	79.84	721	20.16	3576
Reviewer 2	2781	77.77	795	22.23	3576
Reviewer 3	2817	78.78	759	21.22	3576

Table 32: Error indicator per reviewer

When it comes to segments, the reviewers show similar patterns globally despite having different classification or errors per translator, as we saw above. Reviewer 2 changed more segments even though she was the fastest reviewer, and Reviewers 1 and 3 have similar figures, although Reviewer 1 found the fewest segments with errors. For a complete list of segments with errors percentages per translator, see Appendix H.

5.4. Error count

There are more No match segments containing errors, and the similarities between Fuzzy and MT matches are high. It would be interesting to examine the number of errors per translator and category to see if there are more errors in the No match categories. We are using the Global error database here since we are focusing on errors per translator and category and not on segments. Table 33 shows the final number of errors per translator, according to Match category, and the total number of errors. The errors per translator are the sum of the errors from the three reviewers. The table is sorted according to ascending total errors. Totals are highlighted in bold.

Translator	Fuzzy match	MT match	No match	Totals
Translator 15	11	9	24	44
Translator 20	10	16	19	45
Translator 08	15	20	13	48
Translator 12	17	18	15	50
Translator 06	13	20	23	56
Translator 23	14	14	29	57
Translator 05	10	19	29	58
Translator 11	19	23	16	58
Translator 17	13	15	30	58
Translator 16	26	16	21	63
Translator 02	14	15	36	65
Translator 09	26	11	36	73
Translator 24	23	22	31	76
Translator 21	18	23	39	80
Translator 14	16	20	51	87
Translator 01	29	23	37	89
Translator 04	21	22	53	96
Translator 22	16	42	40	98
Translator 07	37	24	44	105
Translator 03	27	41	72	140
Translator 19	45	37	64	146
Translator 18	38	43	69	150
Translator 13	32	53	72	157
Translator 10	52	30	85	167
Totals	542	576	948	2066

Table 33: Number of errors per Match category and translator

Table 33 shows that all segment categories contain errors and that all translators have errors in all categories. There are a total of 2,066 errors in the final texts (this is the aggregated total from the three reviewers). A total of 948 errors are found in the No match segments, 576 in the MT match and 542 in the Fuzzy match. We nevertheless see that in 19 cases there are more errors in the No match segments than in the other two categories. In four cases, there are more errors in MT (Translators 8, 12, 11 and 22); in one case (Translator 16) there are more errors in Fuzzy match. For those translators with over 100 errors, the No match category consistently has the highest amount of errors, and there is more variance in those translators with fewer errors. We can also see that in 15 cases there are more errors in MT than in the Fuzzy match category. These results, however, are the total error numbers; they do not take into account the number of words processed in each category, which was somewhat lower for Fuzzy matches. Although it is unlikely that the result would change for the No match category taking into account the volume, it could change for the other two categories. For example, Translators 17, 2, 4 and 18, with lower number of errors in Fuzzy match in absolute values, would present fewer errors in MT match if the volume of words were factored in, that is, they would present a lower error rate per word in MT match.

Translator 15 has the lowest number of errors, followed closely by Translator 20. Translator 10 has the highest number of errors, followed by Translator 13. Note that in section 4.2, Translator 13 was in fact the fastest, which might indicate that this translator processed the segments very fast but she did not fully post-edit or edit the segment. However, Translator 3, the slowest translator, has 140 errors and is placed not too far from Translator 13 in terms of errors. This might suggest that spending more time on the texts does not necessarily mean a lower number of errors. As we have observed repeatedly in this study, there are striking differences between translators when processing the segments. It is also interesting to turn now to section 3.5 and see that the type of queries these two translators posed at the beginning of the task and their difficulties as initial queries might give the project manager indications on the quality the translators would produce. The next section will examine correlations between speed and number of errors. Let us now look more in depth at the results of errors per category.

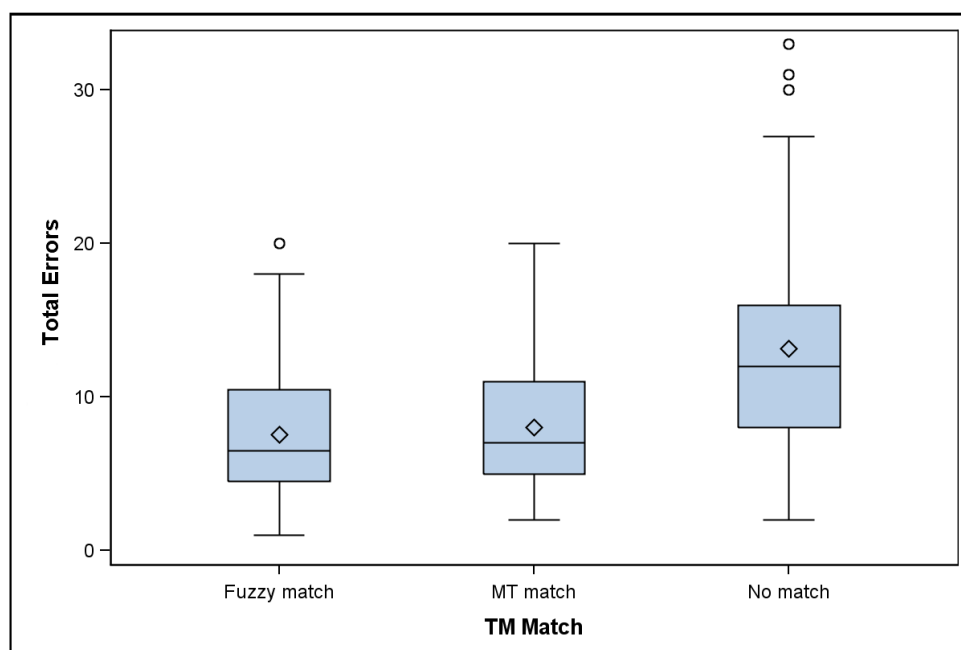


Figure 36: Total errors according to Match category

Figure 36 shows more errors in No match, and similar number of errors in the Fuzzy and MT match categories, although slightly lower for Fuzzy, which suggests that in this case the proposed translations helped participants to produce better quality and at a faster speed. The No match category shows more dispersion, that is, there are more differences between translators. The outliers indicate that one translator had more than 30 errors while others had fewer than 10. The ranges are smaller in the Fuzzy and MT match categories, from above 10 to around 20 errors. Table 34 shows the statistical descriptive data of the sample.

Match	N	Mean	Median	SD	Min	Max
Fuzzy match	72	7.53	6.50	4.11	1.00	20.00
MT match	72	8.00	7.00	4.15	2.00	20.00
No match	72	13.17	12.00	7.22	2.00	33.00

Match	N	Min	Q 1	Median	Q 3	Max	Range	Q Range
Fuzzy match	72	1.00	4.50	6.50	10.50	20.00	19.00	6.00
MT match	72	2.00	5.00	7.00	11.00	20.00	18.00	6.00
No match	72	2.00	8.00	12.00	16.00	33.00	31.00	8.00

Table 34: Descriptive analysis of total errors

Here again the figures for Fuzzy and MT match are very close, and those for No match stand out with higher errors, higher deviations and wider ranges. Translator 10 shows the maximum value, with 33 errors in No match (according to Reviewer 3), and Translator 15 shows the minimum value with 1 error in Fuzzy match (also according to

Reviewer 3). The mean and median values are very close between MT match and Fuzzy match and higher for No match. In general, errors in Fuzzy and MT match are more homogenous than in No match, and this might suggest that the translators that make most errors will make more working on their own (No match) than when they are using the proposed translation (Fuzzy or MT match), while translators with fewer errors have less of a difference between categories. For a full table of errors per translator and reviewer see Appendix G.

But are these differences significant? A Poisson regression model with repeated measures taking as a variable *Total errors* and an offset equal to the *text length* was setup and statistically significant differences were found between the three types of translations ($F=52.48$ and $p<0.0001$). There are differences between the three Match categories, but are results significantly different for MT and Fuzzy matches? From this model, the estimated means for the *logarithm of Total errors /segment length* were calculated according to the Match category. To better interpret the results, Table 35 presents the corresponding estimates expressed in the number of errors per segment length depending on the Match category and considering the results in the original text (length Fuzzy match=618 words, MT match = 757 words, No match = 749 words).

Match	Mean	SD	Lower	Upper
Fuzzy match	7.06	0.45	6.23	8.00
MT match	7.50	0.47	6.63	8.49
No match	12.35	0.70	11.03	13.82

Table 35: Estimated mean of errors in original text

The estimated mean of error in Fuzzy match is 7.06 with a confidence interval of (6.23, 8) in MT match, 7.5 with a confidence interval of (6.63, 8.49), and finally in No match 12.35 with a confidence interval of (11.03, 13.82). It can clearly be seen that the No match is significantly different whereas the other two categories, Fuzzy and MT match, are not.

Our second hypothesis stated that the final quality of the revised target segments translated using MT technology is higher, if measured in errors, than the final quality of revised Fuzzy match segments and lower than the final quality of revised No match segments. We now see that the hypothesis is not validated in this study, since the number of errors in the No match category is significantly higher than in the other two categories and there are no statistically significant differences between Fuzzy and MT match. Translators made more errors when translating without a proposal and made

very similar amounts of errors when editing text from machine translation or translation memories fuzzy match segments from the 85-94 percent range. This is quite different from our previous findings (Guerberof 2008) and is to a certain extent surprising since our hypothesis in the pilot project set out to test what we can see in this project: that the number of errors would not be affected by using MT. It is difficult to establish the reasons for this, since the two projects are quite different in terms of engine used, volume of words, translation memories used, the text itself, the number of translators, the translators themselves, the reviewers, the instructions received and the statistical analysis. We can only hypothesize that the results now are more accurate due to the fact that the volume of words to translate and number of translators were higher and therefore it is closer to a real-life scenario, not to mention the fact that the instructions sent out to the translators were written with the experience drawn from the pilot project. This time around, we felt the need to clearly indicate that the source text reference must be followed and a certain quality had to be achieved. And again, this may be a reflection of the quality of the original translation memories and glossary. With regards to other studies, Tatsumi (2010) did not assess the final quality of the post-edited text, and De Almeida and O'Brien (2010) assess the changes made in the post-edited text but not the quality of the resulting post-edited text. Plitt and Masselot (2010) arrive at a similar conclusion in their study as they establish that the sentences with errors in the translation sample were higher than those in the post-edited sentences for German, French, Italian and Spanish. Fiederer and O'Brien (2009) find that ten raters judged the post-edited 30 sentences into German to be of higher clarity and accuracy, while the translations were judged to be of better style. García (2010) in a project from English into Chinese using Google Translator Toolkit (GTT) found no significant differences between translations using MT and others described as "entirely human", and although reviewers gave different ratings the overall assessment was favorable to MT-seeded texts. There are a number of commercial pilots that establish no difference in quality, including IBM (Roukos et al. 2011) and Sybase (Bier and Herranz 2011). In a recent Translation Automation Conference (TAUS 2011) a PayPal representative stated that "human quality was not good enough for PayPal" (Dove et al. 2011) and she explained in detail the reasons for this, most importantly because MT clarified the meaning with heavily tagged segments, MT did not miss trailing spaces and it performed better in consistency. In a similar study to this one, De Sutter and Depraetere (2012: 13) conclude that "productivity increases without jeopardizing final translation quality" and

that the quality of the post-edited translation and human translation is similar. Our results are therefore not that dissonant in relation to what is being discussed in similar studies and at conferences about this same topic.

5.5. Error classification

In Fiederer and O'Brien (2009), raters judged human translations better in terms of style. We were also interested in seeing the error behavior in terms of type of errors and Match category. Errors have been analyzed and distributed according to the LISA standard to see if the typology of errors varies depending on the type of proposed text (Fuzzy or MT match) or without any translation proposal. The data according to error type are shown in Table 36.

Type of error	No match	MT match	Fuzzy match	Totals	% No match	% MT match	% Fuzzy match	% Total
Mistranslation	59	99	43	201	29%	49%	21%	10%
Accuracy	104	82	125	311	33%	26%	40%	15%
Terminology	227	106	203	536	42%	20%	38%	26%
Language	364	190	142	696	52%	27%	20%	34%
Consistency	1	0	1	2	50%	0%	50%	0%
Country	0	0	6	6	0%	0%	100%	0%
Format	10	16	6	32	31%	50%	19%	2%
Style	183	83	16	282	65%	29%	6%	14%
Totals	948	576	542	2066	46%	28%	26%	100%

Table 36: Number and percentage of errors per type of error

There are 696 Language errors, representing 34 percent of the total number of errors, and 364 of them, that is, 52 percent of all the errors are found in the No match segments. There are 536 Terminology errors, representing 26 percent of the total, and 227 of them, that is, 42 percent of all errors are found, are in the No match segments also. There are 311 Accuracy errors, representing 15 percent of the total errors and 125 of them, 40 percent, are in the Fuzzy match segments and 33 percent in the No match. Of the 201 Mistranslation errors, which are 10 percent in total, 99 are in MT matches, representing 49 percent of the total for this category. There are 282 Style errors in total, representing 14 percent of all errors, and 183 of them are found in the No match segments. If the overall figures are considered, in 75 percent of cases, No match comes highest in percentage of errors. It is interesting to note that although the same translation memory was used for the Fuzzy match category and for training the engine, the number of Terminology errors in MT match is lower. Perhaps, translators checked the glossary more often when they encountered segments with blatant mistakes. Further, the Fuzzy

match category has a low number of Style errors and this could be explained by the quality of the original TM. Figure 37 and Figure 38 below illustrate the number of errors per category and type.

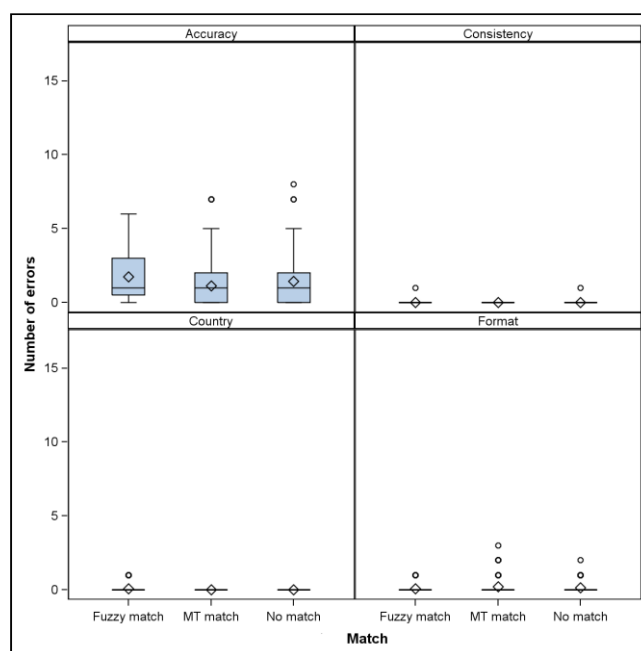


Figure 37: Classification of errors

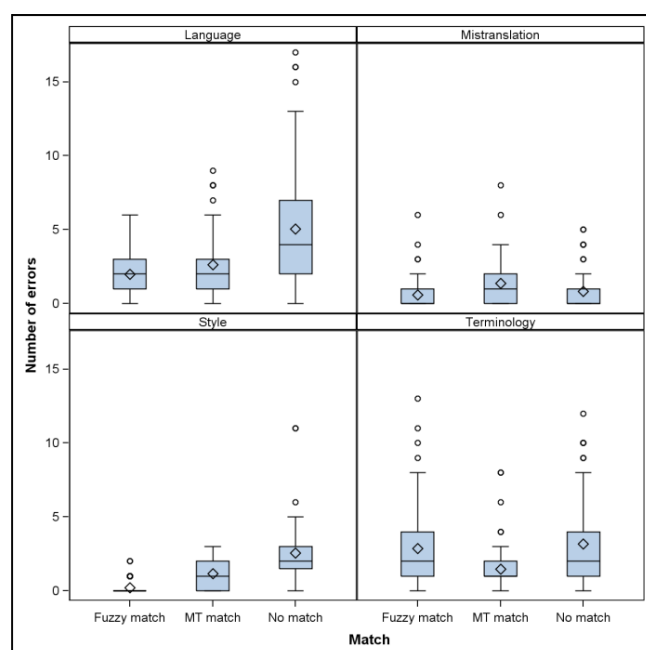


Figure 38: Classification of errors II

These findings are different from those of Fiederer and O'Brien (2009), where raters judged Style to be better in human translated segments, whereas in our study the number of Style errors is higher in the No match segments. However, the instructions

given to the evaluators in both studies differed, and this could account for the different ratings. In Fiederer and O'Brien (2009) raters were asked to judge translations according to Clarity, Accuracy and Style, while this study presented the raters with more options in terms of type of errors and slightly different definitions for Style. Style here included adherence to instructions and consistency throughout the texts, while in Fiederer and O'Brien's study, it referred primarily with tone and natural flow of language. One possible explanation for this number of errors in the No match segments could be that only one translator processed this version (there was no revision) while the other two categories, one way or another, have had several revisions (over time because they come from translation memories), and this points to the quality of the original translation memory and customized use of the engine. We also observe that the translators seem to have checked the glossary less when they were processing No match segments, perhaps trusting their own terminological knowledge, but which seems to have resulted in proportionally more terminological errors. These results are very different from the ones we saw in our pilot project (Guerberof 2008), where the majority of errors were in the Fuzzy matches, followed by MT matches and finally by No match, and the majority of errors were in Accuracy (precisely because translators had not changed the TM segments to match the source text), followed by Language and then Terminology. Apart from all the reasons mentioned in other sections about the differences between the two projects, it is important to highlight that in this case there were three reviewers, while in the pilot project (Guerberof 2008) we were the "judges" of the translation and this could denote a certain revision style. Since there are three reviewers in this case, let us look at the similarities in categorization of the errors.

Type of error	No match	MT match	Fuzzy match	Totals	% No match	% MT match	% Fuzzy match	% Total
Mistranslation	0	32	14	46	0%	70%	30%	7%
Accuracy	30	11	45	86	35%	13%	52%	13%
Terminology	89	32	73	194	46%	16%	38%	29%
Language	124	64	43	231	54%	28%	19%	35%
Consistency	1	0	1	2	50%	0%	50%	0%
Country	0	0	0	0	0%	0%	0%	0%
Format	3	11	3	17	18%	65%	18%	3%
Style	53	21	8	82	65%	26%	10%	12%
Totals	309	171	187	667	46%	26%	28%	100%

Table 37: Reviewer 1 number and percent of errors per type of error

Table 37 shows results for Reviewer 1. These are not that different from the global results: 46 percent of all errors are No match, exactly as before. The difference

here is that the total results for Fuzzy and MT seem to be inverted, with 28 and 26 percent respectively. Regarding categories, No match has still more errors in the categories Language, Style and Terminology. MT has more Mistranslation and Format errors. Fuzzy match has more Accuracy errors. Terminology is low in MT and Style is low in Fuzzy matches as well.

Type of error	No match	MT match	Fuzzy match	Totals	% No match	% MT match	% Fuzzy match	% Total
Mistranslation	1	3	2	6	17%	50%	33%	1%
Accuracy	66	61	61	188	35%	32%	32%	28%
Terminology	63	40	69	172	37%	23%	40%	26%
Language	104	51	61	216	48%	24%	28%	32%
Consistency	0	0	0	0	0%	0%	0%	0%
Country	0	0	6	6	0%	0%	100%	1%
Format	0	0	0	0	0%	0%	0%	0%
Style	53	18	7	78	68%	23%	9%	12%
Totals	287	173	206	666	43%	26%	31%	100%

Table 38: Reviewer 2 number and % of errors per type of error

Table 38 shows results for Reviewer 2. These are still in line with the global results but with slight differences. No match has the majority of errors with 43 percent, followed by Fuzzy matches with 31 percent and lastly by MT matches with 26 percent. The majority of Language, Style and Accuracy errors are placed within the No match category, while in this case Terminology and Country errors are the majority in the Fuzzy matches. MT still has the majority of Mistranslation errors. Style errors are low in Fuzzy matches and Terminology in MT matches.

Type of error	No match	MT match	Fuzzy match	Totals	% No match	% MT match	% Fuzzy match	% Total
Mistranslation	49	64	27	140	35%	46%	19%	19%
Accuracy	8	10	19	37	22%	27%	51%	5%
Terminology	75	34	61	170	44%	20%	36%	24%
Language	136	75	38	249	55%	30%	15%	34%
Consistency	0	0	0	0	0%	0%	0%	0%
Country	0	0	0	0	0%	0%	0%	0%
Format	7	5	3	15	47%	33%	20%	2%
Style	77	44	1	122	63%	36%	1%	17%
Totals	352	232	149	733	48%	32%	20%	100%

Table 39: Reviewer 3 number and percent of errors per type of error

Table 39 shows the results for Reviewer 3. They follow the global trend, albeit with slight changes. No match is still the category with most errors (48 percent), followed by MT matches and lastly Fuzzy matches. No match predominates in Style, Language, Terminology and Format. Fuzzy match has the majority in Accuracy, followed by Terminology and Format and MT match in Mistranslation, followed by Style and Format.

Although there are differences in the classification of errors by the reviewers, they all agree that the No match category has more errors overall and that Language and Style were problematic areas in this particular category. In the case of Fuzzy matches, Accuracy seems to present problems, and for MT matches, Mistranslations. This seems quite logical: in the case of No matches, these strings have never been reviewed (this was, in fact, the first time they were translated) while in the case of translation memories and machine translation, the segments had been “extracted” from the original translation memory. In the case of Fuzzy matches, the main changes to be made in the 85-94 range are related to single words and therefore if this change is missed, there will be more probabilities of Accuracy errors. In the case of MT, the engine might produce an output that is different in meaning, which would need to be completely rearranged or rewritten, thus causing Mistranslation errors. Overall, however, the No match category is not exempt from this type of error. Another interesting aspect is that Fuzzy match has a low percentage of Style errors, indicating the high quality of the original TM, and MT has a low percentage of Terminology errors. We are unsure about the reasons for this, since the same TM was used to train the engine. It could be that translators when correcting blatant errors in MT segments consulted the glossary more frequently. Analyzing the differences between reviewers and the classification of errors is not within the scope of the present study. We feel, however, that there is good reason to analyze this sample further and seek interesting conclusions.

5.6. Errors vs. processing speed

The increase in productivity using different aids, TM or MT, cannot be analyzed in isolation without considering the final quality obtained when using these aids. In other words, there is no advantage in using a pre-translated text and increasing productivity by a certain percentage if as a result more time is needed to review the text to obtain a quality similar to a human translation without any aid. Our sub-hypothesis claimed that *translators with higher overall processing speeds when using MT or TM technology will have fewer errors than those with lower processing speeds*. Therefore, we need to compare the translators’ processing speeds in relation to the number of errors in the final target texts. Table 40 shows the total processing speed of the 24 participants, resulting from looking at the mean value in processing speed, sorted from the highest to

the lowest (words per minute) and the number of errors. Translators with over 100 errors are highlighted in bold.

Translator	Mean processing speed	Global error count
Translator 3	9.39	140
Translator 11	9.54	58
Translator 4	10.23	96
Translator 14	11.27	87
Translator 12	12.24	50
Translator 2	12.86	65
Translator 9	13.21	73
Translator 17	14.64	58
Translator 22	15.11	98
Translator 6	15.73	56
Translator 16	15.77	63
Translator 24	16.12	76
Translator 21	17.83	80
Translator 20	20.05	45
Translator 7	20.88	105
Translator 8	21.43	48
Translator 1	21.85	89
Translator 23	22.03	57
Translator 10	22.3	167
Translator 18	22.6	150
Translator 19	25.1	146
Translator 15	25.79	44
Translator 5	26.11	58
Translator 13	31.59	157

Table 40: Translators' processing speed versus number of errors

Translator 13, with the fastest time, in words per minute, has 157 errors, while Translator 3, with the slowest time, has 140 errors. These two translators also had difficulties at the beginning of the project that might indicate future quality issues (see section 3.5). In these two minimum and maximum cases, speed does not seem to have played an important role in the final error count. On the other hand, translators with fewer errors (Translators 15, 5 and 23) have a high processing speed if compared to most of the others, but other translators (Translators 11 or 12) that also have a low number of errors have a low processing speed. Translators 4 or 22, with errors close to 100, are not among those with higher processing speeds. A priori, by just looking at this overall data, it is difficult to find support for our second hypothesis. The mean value tells us the global average speed of each translator and not necessarily the speed per category or segment. To see the differences more accurately, we decided to group translators and compare their speeds versus the number of errors. We had the classification into "Speed groups" done according to the No match category.

- Group 1: Less than 10 words per minute. Translators: 2, 3, 4, 6, 9, 11, 12, 14 and 22.

- Group 2: From 10 (included) to 15 words per minute. Translators: 8, 16, 17, 18, 21 and 24.
- Group 3: 15 or more words per minute. Translators: 1, 5, 7, 10, 13, 15, 19, 20 and 23

This classification was used and the No match category was set as the baseline speed for translators. Figure 39 shows these results.

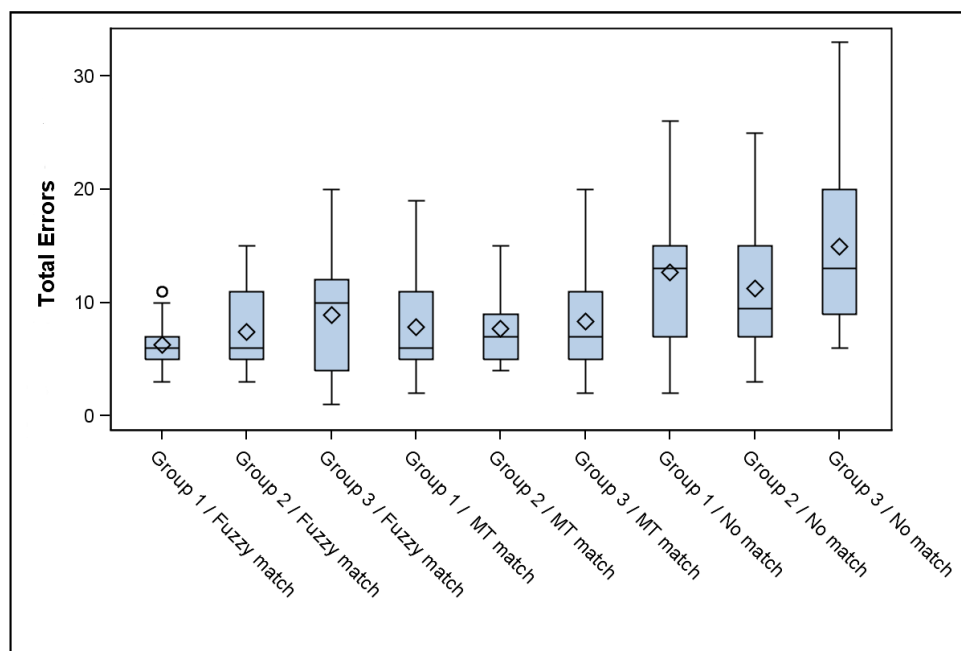


Figure 39: Speed group and number of errors

Group 3 has the highest mean processing speed in the No match category. However, the number of errors does not appear to be lower in any of the three categories, as our sub-hypothesis claims, but higher than in Groups 2 and 1, although the ranges are quite wide. Group 1 performs better (fewer errors) in the Fuzzy match category and Group 2 in the MT match category, but the differences are not pronounced. The descriptive data per category illustrate these observations better.

Speed group	Rev	N	Mean	Median	SD	Min	Max
Group 1	Rev 1	9	6.44	6.00	1.81	4.00	10.00
	Rev 2	9	7.56	7.00	2.07	5.00	11.00
	Rev 3	9	4.78	4.00	2.05	3.00	9.00
Group 2	Rev 1	6	8.00	7.50	4.56	3.00	15.00
	Rev 2	6	8.17	7.00	3.06	5.00	12.00
	Rev 3	6	6.00	5.00	2.53	4.00	11.00
Group 3	Rev 1	9	9.00	10.00	5.68	2.00	18.00
	Rev 2	9	9.89	11.00	5.04	4.00	16.00
	Rev 3	9	7.78	8.00	6.06	1.00	20.00

Table 41: Errors vs. Speed groups in Fuzzy match

If the mean value of errors is to be considered in the Fuzzy match category the value is higher in Group 3, then in Group 2 and finally in Group 1. However, the minimum and maximum values in all three categories indicate that there are translators that made very few errors in Group 3, while others in the same group made more resulting in a higher mean value. The deviation is higher in the Group 3, the fastest. When the overall number of errors is examined per translator, some of the translators with more errors are in this group: Translators 7, 10, 13, 19, but also those with fewer errors: Translators 15, 20 and 5. It is understandable that we see extreme values, since the deviations are very pronounced in this group. The maximum values correspond to (according to Reviewer 3) Translator 10 (20), (Reviewer 2) Translator 10 (16) and (Reviewer 1) Translator 19 (18), and the minimum to (according to Reviewer 3) Translator 15 (1), (Reviewer 2) Translators 5 and 20 (4) and (Reviewer 1) Translator 20 (2). The results between Group 1 and 2 are very similar. Notwithstanding, Group 1 has slightly lower results.

Speed group	Rev	N	Mean	Median	SD	Min	Max
Group 1	Rev 1	9	6.78	6.00	3.60	2.00	13.00
	Rev 2	9	7.22	6.00	3.15	3.00	13.00
	Rev 3	9	9.56	9.00	4.77	5.00	19.00
Group 2	Rev 1	6	7.00	7.00	3.29	4.00	13.00
	Rev 2	6	7.00	6.00	4.00	4.00	15.00
	Rev 3	6	9.17	8.00	3.71	5.00	15.00
Group 3	Rev 1	9	7.56	6.00	4.28	3.00	14.00
	Rev 2	9	7.33	6.00	5.05	2.00	19.00
	Rev 3	9	10.11	10.00	4.96	4.00	20.00

Table 42: Errors vs. Speed groups in MT match

Table 42 shows the results for the MT match category. Here the mean values are more homogenous between the three groups, only slightly higher in Group 3. Still, the ranges are wide, meaning that translators performed quite diversely within each group. We can see this in the minimum and maximum values.

Speed group	Rev	N	Mean	Median	SD	Min	Max
Group 1	Rev 1	9	11.67	13.00	6.91	2.00	25.00
	Rev 2	9	11.78	11.00	5.19	6.00	21.00
	Rev 3	9	14.56	13.00	7.65	4.00	26.00
Group 2	Rev 1	6	12.83	12.50	6.97	5.00	24.00
	Rev 2	6	9.83	8.50	5.64	3.00	20.00
	Rev 3	6	11.17	9.50	7.31	5.00	25.00
Group 3	Rev 1	9	14.11	13.00	7.41	6.00	31.00
	Rev 2	9	13.56	9.00	7.52	6.00	27.00
	Rev 3	9	17.11	14.00	9.88	6.00	33.00

Table 43: Errors vs. Speed groups in No match

Table 43 shows the number of errors per speed group for the No match category. Although the number of errors is higher overall in the No match category, as seen above, there is only a very slight difference in Group 3. However, the three groups behave in a similar way. After looking at the speed groups, we do not observe real differences in errors in relation to speed. A Poisson regression model with repeated measures was applied taking *Total errors* as the response variable and an offset equal to the text length. No statistically significant differences were observed between the Speed groups or in the interaction between Match category and Speed groups.

However, this was not the overall speed per translator - these groups were distributed according to the No match speed. The question then was, is the speed different for translators in other categories? We decided then to group translators according the speed in each category in the following manner:

- Group 1 ("Slow group"): fewer than 10 words per minute in No match; fewer than 15 words per minute in Fuzzy and MT match.
- Group 2 ("Intermediate Group"): between 10 and 15 words per minute in No match; and 15 and 20 words per minute in Fuzzy and MT match.
- Group 3 ("Fast group"): more or 15 words per minute in No match and 20 or more words per minute in Fuzzy and MT match.

Table 44 shows this distribution per translator.

Translator	Fuzzy match	MT match	No match
Translator 01	3	3	3
Translator 02	1	2	1
Translator 03	1	1	1
Translator 04	1	1	1
Translator 05	3	3	3
Translator 06	2	3	1
Translator 07	3	3	3
Translator 08	3	3	2
Translator 09	2	1	1
Translator 10	3	3	3
Translator 11	1	1	1
Translator 12	1	1	1
Translator 13	3	3	3
Translator 14	1	1	1
Translator 15	3	3	3
Translator 16	2	2	2
Translator 17	2	2	2
Translator 18	3	3	2
Translator 19	3	3	3
Translator 20	3	3	3
Translator 21	3	3	2
Translator 22	2	2	1
Translator 23	3	3	3
Translator 24	2	2	2

Table 44: Translators according to three Speed groups

There are seven translators that change groups (marked in bold), that is, their speed group with MT and Fuzzy match is different from their speed group considering only the No match category. Most of the translators that changed group were previously in Speed groups 1 and 2 considering the No match category. The changes are small, so we do not expect to see significant consequences overall. A Poisson regression model was adjusted for Fuzzy matches with repeated measures taking *Total errors* as the response variable and as offset of the text length (because the word counts for each Match categories are different) and there were no statistically significant differences for the different speed groups. The same thing occurred for MT match and for No match.

We can therefore conclude that speed during the assignment did not affect the final quality obtained in the task according to the three reviewers. Quality, measured in number of errors, could be related to the level of experience of translators (we will examine that in the following section) but certainly not to the word processing speed in our project. We can observe, however, that in some instances fast translators made fewer errors than other slower translators and this could explain our observations in our previous project (Guerberof 2008) where we had eight translators to compare against each other. This is in line with the study Künzli (2007) looked at the revisions of a legal text done by ten translators and the results indicated that more time did not mean a high-

quality text. Lorenzo (2002), in an experiment to test revision competence among students, also finds that longer revision time does not correlate with fewer errors.

5.7. Overcorrections

As explained in section 3.4.3.7, the reviewers were instructed to mark overcorrections, that is, to mark the edits or post-edits that translators had made but that went beyond what was needed, and, at the same time, they were instructed not to count them as errors.

These are the results obtained:

Translators	Speed group	Reviewer 1	Reviewer 2	Reviewer 3	Total
TR01	3	0	2	3	5
TR02	1	1	0	2	3
TR03	1	0	0	3	3
TR04	1	0	0	3	3
TR05	3	0	0	0	0
TR06	2	0	0	3	3
TR07	3	0	0	1	1
TR08	3	0	1	8	9
TR09	2	0	0	4	4
TR10	3	0	0	6	6
TR11	1	0	0	2	2
TR12	1	0	0	3	3
TR13	3	0	0	4	4
TR14	1	0	0	3	3
TR15	3	0	0	0	0
TR16	2	0	0	0	0
TR17	2	0	0	4	4
TR18	3	0	0	10	10
TR19	3	1	0	10	11
TR20	3	0	0	2	2
TR21	3	0	0	2	2
TR22	2	0	1	4	5
TR23	3	1	0	2	3
TR24	2	0	1	3	4
Total		3	5	82	90

Table 45: Overcorrections and Speed groups in Fuzzy match

Translators	Speed group	Reviewer 1	Reviewer 2	Reviewer 3	Total
TR01	3	0	1	5	6
TR02	2	0	0	3	3
TR03	1	0	0	4	4

Translators	Speed group	Reviewer 1	Reviewer 2	Reviewer 3	Total
TR04	1	0	0	2	2
TR05	3	0	0	1	1
TR06	3	0	0	1	1
TR07	3	0	0	1	1
TR08	3	0	1	1	2
TR09	1	0	0	7	7
TR10	3	0	0	11	11
TR11	1	0	0	3	3
TR12	1	0	0	2	2
TR13	3	0	0	0	0
TR14	1	0	0	1	1
TR15	3	0	0	2	2
TR16	2	0	0	0	0
TR17	2	0	0	0	0
TR18	3	0	0	8	8
TR19	3	1	1	7	9
TR20	3	0	0	0	0
TR21	3	0	0	1	1
TR22	2	0	0	3	3
TR23	3	0	0	1	1
TR24	2	0	1	0	1
Total		1	5	64	70

Table 46: Overcorrections and Speed group in MT match

Translators	Overcorrections	Errors
TR01	11	89
TR02	6	65
TR03	7	140
TR04	5	96
TR05	1	58
TR06	4	56
TR07	2	105
TR08	11	48
TR09	11	73
TR10	17	167
TR11	5	58
TR12	5	50
TR13	4	157
TR14	4	87
TR15	2	44
TR16	0	63
TR17	4	58
TR18	18	150
TR19	20	146

Translators	Overcorrections	Errors
TR20	2	45
TR21	3	80
TR22	8	98
TR23	4	57
TR24	5	76
Total	160	2066

Table 47: Overcorrections vs. global errors

As it can be seen in Table 45 and Table 46, Reviewers 1 and 2 marked very few overcorrections, 4 and 10 respectively while Reviewer 3 marked 146 in total. This might reflect the fact that the reviewers were informed that overcorrections were not to be marked as errors, hence, Reviewers 1 and 2 did not think they were important, while Reviewer 3 spent more time marking this (this Reviewer also identified more errors overall). Perhaps, the instructions should have been clearer on what an overcorrection was and how to classify it (see Appendix C). The only conclusion we can draw from the figures is that Reviewers 1 and 2 found very few overcorrections overall, and Reviewer 3 found more overcorrections. Reviewers 1 and 3 found more overcorrections in Fuzzy match than in MT match, although Reviewer 1 found very few overall, and Reviewer 2 found an equal amount of corrections in both categories. If we look closely at the translators that had more overcorrections marked in bold in Table 47 (Translators 1, 8, 10, 18 and 19), we find that they are also among the translators with more errors globally, except for Translator 8. If we look closely at the translators that had fewer errors, we find that they are also among the translators with fewer overcorrections (Translators 5, 6, 11, 12, 15, 17, 20 and 23). However, in some cases the opposite is also true: some translators with greater number of errors have fewer overcorrections, and others with fewer errors have a greater number of overcorrections (Translators 3, 4, 13 and 8). If we look at the Speed group and number of overcorrections per Match category, we can see that the translators with more overcorrections (marked in bold in Table 45 and Table 46) had a higher processing speed (Speed group 3). This is interesting as it might indicate that perhaps these translators were working too fast and not being sufficiently thorough. However, there are other translators in Speed group 3 that have fewer overcorrections. Although the aim of this study was not to investigate the concept of preferential changes, we believe that this is a topic that would need to be studied further in order to develop instructions or training materials for post-editing.

5.8. Conclusions on quality

After an analysis of the data taken from the errors found during the review process, we have observed that reviewers presented differences when correcting the translations from the 24 translators both in time, words reviewed per minute, and in number of errors. However, when comparing the reviewers against each other, we have found that they tended to agree on the number of No match errors but their agreement in Fuzzy and MT match was either weak or there was no agreement. However, the reviewers agreed that most segments (78.79 percent) did not contain errors and that the No match category had a higher percentage of segments with errors overall. The percentage of segments with errors in Fuzzy and MT match had almost identical values, 18.17 and 18.11 percent respectively. We have also found that the odds of making a mistake in No match are 1.75 times the odds of making a mistake in the Fuzzy and the MT match categories.

When we looked at the number of errors per translator we found significant differences between the three categories of matches but these differences were not significant between Fuzzy and MT matches. This indicates that the proposed text helped translators to produce better quality, if we consider that the lower the number of errors, the higher the final quality of the text. Our second hypothesis, however, which claimed that *the final quality of the revised target segments translated using MT technology is higher, if measured in errors, than the final quality of revised Fuzzy match segments and lower than the final quality of revised No match segments* is not validated in this study, since the number of errors in the No match category is significantly higher than the one in the other two categories. Translators made more errors when translating without a proposal and made a very similar number of errors when editing text from MT or TM segments from the 85-94% range. These results are in line with other studies, e.g. Plitt and Masselot (2010), García (2010), Fiederer and O'Brien (2009), and De Sutter and Depraetere (2012).

Regarding the type of errors found in the study, 48 percent of all errors are in the No match segments, 28 percent in the MT match category and 26 percent in the Fuzzy match category if we consider absolute number of errors (that is, without considering the number of words processed). The No match category has the majority of Language, Terminology and Style errors, while the Fuzzy match category has the majority of Accuracy errors and the MT match category, the Mistranslation errors. Although there

are differences in the classification of errors according to each of the reviewers, they all agree that the No match category has more errors overall and that Language and Style were problematic areas in this particular category. In the case of Fuzzy matches, Accuracy seems to present problems, and for MT matches, Mistranslations. However, Terminology errors are low in MT matches and Style errors are low in Fuzzy matches.

With regard to overcorrections, Reviewers 1 and 2 found very few overall, and Reviewer 3 found significantly more in some translators. Reviewers 1 and 3 found more overcorrections in Fuzzy match than in MT match, and Reviewer 2 found an equal amount of corrections in both categories.

We have tested our sub-hypothesis that claims that *translators with higher processing speeds, words per minute, overall processing speed when using MT or TM technology, will have fewer errors than those with lower processing speeds*. It was found that the speed variable does not present statistically significant differences within all the speed groups. We can then conclude that speed during the assignment did not affect the final quality obtained in the task according to the three reviewers. It seems that it was independent of speed. Quality, measured in number of errors, could be related to the level of experience of translators (which we examine in the following chapter) but certainly not to the processing speed in the project. We can observe, however, that in some instances fast translators made fewer errors than slower translators and this could explain our observations in our previous project (Guerberof 2008) where there were only eight translators to obtain data from.

Chapter 6: Translators' experience results

In this chapter, information on the experience of the participants according to their responses in the post-assignment questionnaire will be examined in order to test the third hypothesis of this study, which relates more experience to an increase in speed but not necessarily with an increase or decrease in the number of errors. We will also explain how the participants were grouped into different clusters. The translators' experience will be correlated with speed according to the words per minute that each cluster processed in the different Match categories (Fuzzy, MT and No match). Any peculiarities in the interaction between the clusters, match categories and speed will be observed. Secondly, we will look at the behavior of these clusters in relation to the errors marked by the three reviewers in each match category. Finally, conclusions will be drawn in relation to experience, speed and number of errors.

6.1. "About your experience"

The first question that comes to mind when starting in this area of the research is "What does experience mean?". We are aware that the term embraces several aspects of a translator's experience. For the purpose of this study, experience is defined as a combination of years of experience in localization, subject matter, tools knowledge, post-editing, type of tasks performed, estimation of daily throughputs and average typing speed (as explained in section 3.3). The data were obtained from the first section of the questionnaire that was provided to the translators through SurveyMonkey upon completion of the assignment (see Appendix D). In section "About your experience", pages 2 to 4 of this questionnaire, the translators responded to the following questions:

- How long have you been working in the localization industry?
- How long have you been using translation memory tools (such as SDL Trados, Star Transit, Déjà Vu)?
- How long have you been translating business intelligence software (such as SAP, Oracle, Microsoft)?
- Have you done translation work for MicroStrategy (directly or through a LSP) in the last three years?

- How long have you been post-editing raw machine translated (MT) output?
- Please estimate the percentage, on average, that post-editing MT output represents in your work (considering the last three years)
- What tasks does your work involve? (You can choose more than one option).
- Please estimate your average daily throughput when you translate from scratch without any translation aid:
- What is your average typing speed? (Please, provide an estimate in words per minute).

We present a brief overview of their responses in order to understand better the experience of the participants before they are grouped into different clusters. A table with the results will be followed by a brief analysis of the participants' responses.

Answer Options	Response Percent	Response Count
No experience.	0.0%	0
Less than 2 years.	0.0%	0
2 years or more, less than 4 years.	12.5%	3
4 years or more, less than 6 years.	12.5%	3
6 years or more, less than 8 years.	25.0%	6
8 years or more.	50.0%	12

Table 48: Experience in the localization industry

The results indicate that this is an experienced group: 50 percent have more than eight years' experience in the industry (Translators 2, 4, 5, 7, 9, 12, 17, 20, 21, 22, 23 and 24), 25 percent have more than six years' experience and less than eight (Translators 1, 11, 13, 14, 15 and 18), 12.5 have more than four years' experience and less than six (Translators 6, 8 and 16) and another 12.5 have more than two years' experience and less than four (Translators 3, 10 and 19). All translators are in a range that goes from more than two years to eight years or more. Exactly 75 percent had more than six years' experience.

Answer Options	Response Percent	Response Count
Never.	0.0%	0
Less than 2 years.	0.0%	0
2 years or more, less than 4 years.	12.5%	3
4 years or more, less than 6 years.	12.5%	3
6 years or more, less than 8 years.	25.0%	6
8 years or more.	50.0%	12

Table 49: Experience in translation memory tools

The responses to the questions on TMs are almost identical to the ones in the previous question: 50 percent have more than eight years' experience using translation memory tools (Translators 2, 4, 5, 7, 9, 12, 17, 20, 21, 22, 23 and 24), 25 percent have more than six years' experience and less than eight (Translators 1, 6, 13, 14, 15 and 18), 12.5 have more than four years' experience and less than six (Translators 8, 11 and 16) and another 12.5 have more than two years' experience and less than four (Translators 3, 10 and 19). There are only two changes with respect to the previous question. Translator 6 had four years' experience in localization and six in using translation memory tools, probably this participant used these tools in other domains. Translator 11, on the other hand, had more years of experience in localization (six) than in tools (four) - presumably this participant started translating in a word processor without any translation memory tool.

Answer Options	Response Percent	Response Count
Never.	8.3%	2
Less than 2 years.	8.3%	2
2 years or more, less than 4 years.	4.2%	1
4 years or more, less than 6 years.	29.2%	7
6 years or more, less than 8 years.	16.7%	4
8 years or more.	33.3%	8

Table 50: Experience in business intelligence translation

For this question, there was a wider variety of responses: 8.3 percent has never translated business intelligence (Translators 3 and 17), 8.3 percent has less than two years' experience (Translators 10 and 19), 4.2 percent has two years or more and less than four (Translator 8), 29.2 percent has four years or more and less than six (Translators 1, 11, 13, 15, 16, 18 and 21), 16.7 percent has six years or more and less than eight years (Translators 4, 6, 14, and 22) and 33.3 percent has eight years or more (Translators are 2, 5, 7, 9, 12, 20, 23 and 24). The experience is more heterogeneous in this group in relation to the subject matter or domain, but still only four translators have less than two years' experience or none, and 20 have considerable experience in business intelligence software translation.

Answer Options	Response Percent	Response Count
Yes.	8.3%	2
No.	87.5%	21
I don't know.	4.2%	1

Table 51: Translation work for MicroStrategy

During the selection process, the Vendor Management team did not select translators who had experience in MicroStrategy. This was to avoid having translators with a faster translation speed on account of experience with this customer. It was nevertheless interesting to ask this question just in case some translators might have worked for other agencies without the Vendor Management team knowing about it or in case they had worked for this account but the VM team was not aware of it. A total of 21 responded that they had not and two responded that they had: Translator 5 and Translator 19. These two translators performed at a high speed (Group 3), and although Translator 5 had a low number of errors, Translator 19 had quite a high number of errors. Thus, this experience might have helped them but it is not obvious by the data obtained. Translator 6 responded that she did not know. This is understandable if we consider that translators work for a multitude of customers through LSPs and on occasions they work on small pieces. We checked with the Vendor Management team and they confirmed that there were two translators that had worked on this account (Translators 5 and 6) but the volume of work carried out was 10,000 and 20,000 words respectively, which is not considered very high in a localization environment. It is possible that Translator 19 had worked on this account for another customer. Consequently, the vast majority of the participants had no experience with this particular customer, despite having experience in business intelligence software translation.

Answer Options	Response Percent	Response Count
Never.	25.0%	6
Less than 2 years.	29.2%	7
2 years or more, less than 4 years.	25.0%	6
4 years or more, less than 6 years.	8.3%	2
6 years or more, less than 8 years.	4.2%	1
8 years or more.	8.3%	2

Table 52: Experience in post-editing

Twenty-five percent of responses were Never (Translators 3, 6, 16, 17, 21 and 22), 29.2 percent (the highest percentage) have less than two years' experience (Translators 1, 5, 10, 12, 13, 18 and 19), 25 percent have two years or more and less than four (Translators 7, 8, 11, 14, 15 and 23), 8.3 percent have four years or more and less than six (Translators 4 and 20), 4.2 percent have six years or more and less than eight (Translator 24), and 8.3 percent have eight years or more (Translators 2 and 9).

This shows that post-editing is a relatively new task for this group in comparison with their experience in the other areas. Exactly 79.2 percent of the whole group has no

experience or less than four years' experience which is low in comparison to the general experience in localization, tools and business intelligence that was described above.

Answer Options	Response Percent	Response Count
0%	25.0%	6
1% to 25%	66.7%	16
26% to 49%	4.2%	1
50% to 74%	4.2%	1
75% to 90%	0.0%	0
91% to 100%	0.0%	0

Table 53: Estimated post-editing work in the last three years

We wanted to qualify the previous questions as some translators might have certain experience in post-editing but not a lot of work in this specific task on a yearly basis. Exactly 25 percent responded that they had 0 percent work on post-editing (Translators 3, 6, 16, 17, 21 and 22), the ones that responded Never to the previous question. Then 66.7 percent estimate 1 to 25 percent of the work (Translators 1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 15, 18, 19, 20, 23 and 24), 4.2 percent selected 26 to 49 percent (Translator 9) and 4.2 percent placed post-editing in the 50 to 74 percent range (Translator 14). Post-editing still does not represent a high percentage of work for this group of translators.

Tasks	No		Yes		Total
	N	Row Percentage	N	Row Percentage	
Post-editing	9	37.50	15	62.50	24
Translating	1	4.17	23	95.83	24
Revising translations	3	12.50	21	87.50	24
Writing	20	83.33	4	16.67	24
Terminology work	15	62.50	9	37.50	24
Other	19	79.17	5	20.83	24

Table 54: Different tasks at work

A total of 37.5 percent responded that they do not do post-editing as part of their tasks (Translators 2, 3, 6, 13, 15, 16, 17, 21 and 22). These are the translators that responded Never to post-editing. Translators 2, 13 and 15, who had given a low percentage in post-editing work done over the last three years, also responded that they did not do post-editing work. On the other hand, 95.83 percent translated (all except Translator 18 who had a high number of errors despite being on a fast speed group), 87.5 percent revise translators (all except Translators 8, 10 and 19). Translators 10 and 19 are novice translators and it might be logical that they are not revising other translators' work. Translator 8 has more experience but perhaps she has chosen not to review or simply she does not receive this type of assignment. Then 83.3 percent do not

write, the ones that do being Translators 2, 6, 15 and 24. Translators 1, 7, 8, 9, 11, 15, 20, 21 and 24 carry out terminology work (37.5 percent). Other tasks specified are: Project Management (Translator 9), Localization engineering and preparation (of files for localization, presumably) (Translator 15), teaching translation (Translator 17), Linguistic Testing, Linguistic Advising and Linguistic QA (Translator 20) and Application testing and Project Management (Translator 24). It caught our attention that Translator 15, with optimal results in terms of speed and errors, had some engineering experience, as this could indicate that certain technical abilities might help when translating and post-editing. Also Translator 20, also with optimal results, had experience in linguistic quality assurance, and this might indicate higher linguistic skills. The main tasks in this group are translating and revising, while post-editing comes in the third place of the proposed tasks.

Answer Options	Response Percent	Response Count
Less than 2000 words per day.	8.3%	2
Between 2100 and 3000 words per day.	70.8%	17
Between 3100 and 5000 words per day.	20.8%	5
More than 5100 words per day.	0.0%	0
I don't know	0.0%	0

Table 55: Estimated daily throughput when translating from scratch

The majority (70.8 percent) selected the option between 2,100 and 3,000 words per day which is considered a standard metric in the industry and thus is not surprising. On the other hand, 8.3 percent estimated less than 2,000 words per day (Translators 9 and 11). These two translators were in the speed group 1 (the slowest speed group) in our productivity section (see section 4.4). The 20.8 percent estimated between 3,100 and 5,000 words per day (Translators 2, 5, 7, 13 and 17). Three out of this group were in the fastest speed group in our study (5, 7 and 13). However, Translator 2 was in Speed group 1 and 17 in Speed group 2.

Answer Options	Response Percent	Response Count
0-20 words per minute	8.3%	2
21-40 words per minute	16.7%	4
41-60 words per minute	41.7%	10
61-80 words per minute	20.8%	5
More than 81 words per minute	12.5%	3
Comments:		3

Table 56: Estimated typing speed

Exactly 41.67 percent had a typing speed of 41 to 60 words. Six translators were slower than this: 2, 8, 11, 16, 18, and 22. These translators were placed in Speed groups

1 and 2 in our study. Translators 2 and 22 were the slowest in the group when typing and they were also in the Speed group 1. Eight translators were faster than 41 to 60 words per minute: 1, 4, 5, 9, 15, 19, 20 and 23. Translators 9, 15 and 23 declared having the fastest typing speed in the group (but Translators 1 and 9 were in the Speed group 1 in our study). There were also three comments: from Translator 2, *“In fact, no idea, sorry... I have never considered this detail”*; from Translator 17, *“Just an estimation as I haven't taken the time to check!”* and from Translator 18, *“It depends on the language, e.g. words in German are (sic) far more (sic) complicated to type than ones in Spanish or English.”*

All responses suggest that this is a group of experienced professionals with slightly different areas of expertise, although there are three translators with less experience than the remaining twenty-one. They also have considerable experience using tools and some experience in post-editing MT output, although the task represents a low percentage of their work and has not been performed for a very long period of time. Their working speed seems to be in accordance with the industry standard and it is quite homogeneous. It might appear difficult, due to their relative homogeneity, to place these translators in different groups. In the next section, we explain how the translators were grouped into clusters.

6.2. Grouping translators according to their experience

In order to distribute translators into different groups with similar experience, a multiple correspondences analysis was setup (Greenacre 2008). This enables us to represent all the data (responses from the questionnaire by all translators) as rows and columns in a table including active variables (the questions above) and showing illustrative variables (age and sex). These were then graphically represented as dots in a two dimensional map (biplot). Four groups (clusters) were found, with distinctive characteristics. To explain the complete statistical analysis is beyond the scope of this study, but we are presenting a sample of a biplot to illustrate how the multidimensional associations are projected in a two-dimensional map when looking at two factors. The factors are not pre-defined, as we plot the data to see how the different variables are related in order to understand their relation and hence define the groups. In this case we have used ten characteristics that serve to explain most of the data.

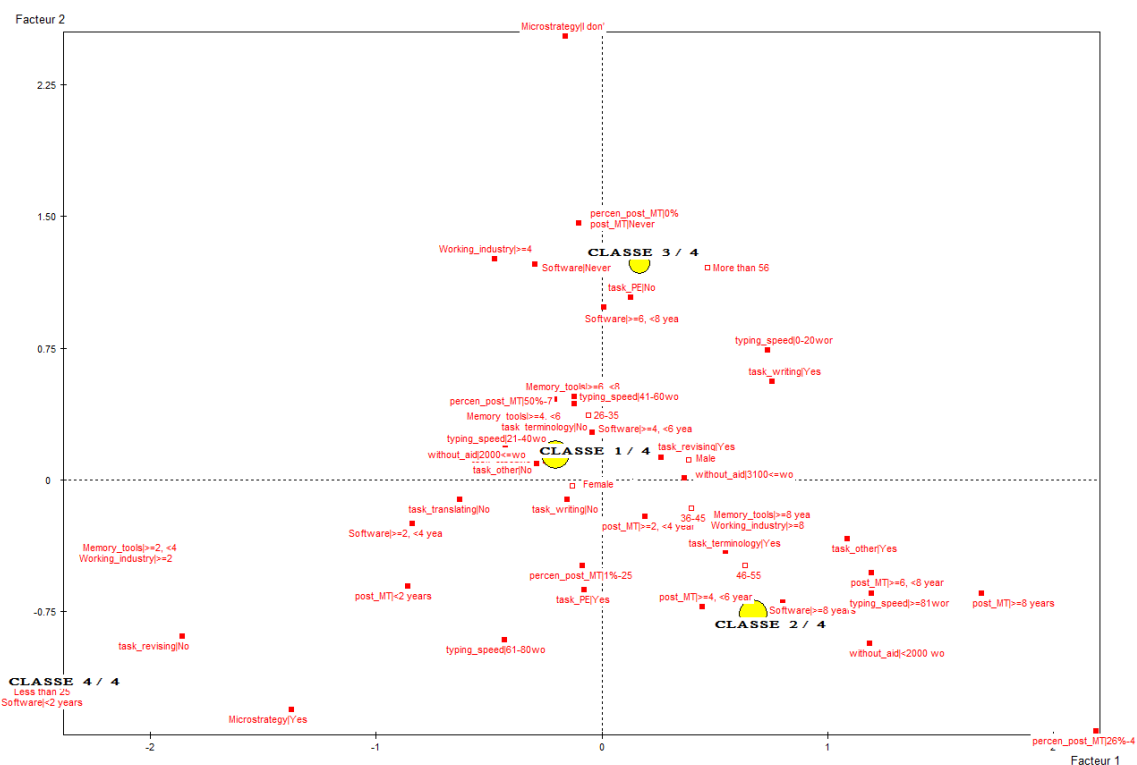


Figure 40: Sample of biplot

The dendrogram below represents the hierarchical structure of the data. This tree specifies the points of union and distance between the different clusters.

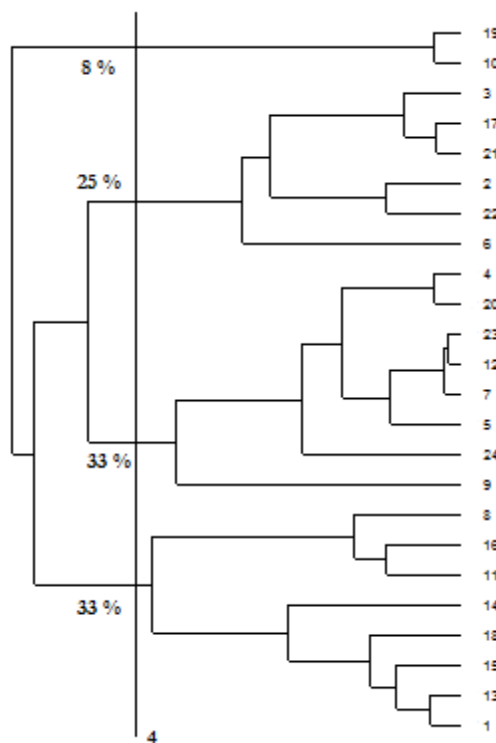


Figure 41: Dendrogram of clusters

For example, Translators 10 and 19 gave very similar responses, so the distance between them is very short (this group represents eight percent of all translators). On the other hand, Translators 8, 16, 11, 14, 18, 15, 13 and 1 presented a greater distance (this group represents 33 percent of all translators), which means that their responses were not as similar as those in the previous cluster but showed more similarities than with the other translators in the whole sample. For example, Translator 13 and 1 show more similarities than Translator 16 and 11 because the first ones joined earlier than the latter ones (see Dendogram). The vertical line (4) indicates the optimum line drawn to define a reasonable number of clusters. If the line is set at a shorter distance, we might obtain 24 different groups; if the line is set at a longer distance, we might obtain only one group of translators.

6.3. Clusters

Translators have been grouped in this particular way because they have common characteristics, as taken from their responses to the active variables explained in the previous section. The description of the groups is as follows:

6.3.1. Features of Cluster 1

This group of translators is characterized by having experience in all the areas queried in the questionnaire, but they have been doing these tasks for a shorter period of time than those in Cluster 2. The translators in this cluster are: 1, 8, 11, 13, 14, 15, 16 and 18. To these questions, the translators responded as follows:

1. *How long have you been working in the localization industry?*

75 percent of these translators (six of them) have between six and eight years' experience in localization, and this represents all the translators with this level of experience within the 24 translators. This percentage is significantly higher than in the global group of translators ($t=3.53$, $p<0.001$).

2. *How long have you been using translation memory tools?*

62.5 percent of the total number of translators in this cluster has between six and eight years' experience in translation memory tools, and this represents 83 percent of the total of translators with this experience in the global group of translators. This percentage is significantly higher than in the global group of translators ($t=2.46$, $p=0.007$).

3. *How long have you been translating business intelligence software?*

75 percent of the translators in this cluster have between four and six years' experience in translating business intelligence, which represents 85 percent of the total number of translators with this experience in the group of 24. This percentage is significantly higher than in the global group of translators ($t=3.01$, $p_value=0.001$).

4. *What is your average typing speed?*

50 percent of the translators in this cluster have a speed ranging from 21 to 60 words per minute, and this represents all the translators with this speed in the group of 24 translators. This percentage is significantly higher than in the global group of translators ($t=2.48$, $p=0.007$).

6.3.2. *Features of Cluster 2*

This group of translators is characterized by having experience in all the aspects queried in the questionnaire. They are the group with the most experience. The translators in this cluster are: 4, 5, 7, 9, 12, 20, 23 and 24. To these questions, the translators responded as follows:

1. *How long have you been working in the localization industry?*

All translators in this cluster have more than eight years' experience in the localization industry, and this represents 66.7 percent of the total of translators with this experience in the whole group of 24. This percentage is significantly higher than in the global group of translators ($t=3.21$, $p=0.001$).

2. *How long have you been using translation memory tools?*

All translators in this cluster have more than eight years' experience using translation memories, and this represents 66.7 percent of the total of translators with experience in the whole group of 24 translators. This percentage is significantly higher than in the global group of translators ($t=3.21$, $p=0.001$).

3. *How long have you been translating business intelligence software?*

87.5 percent of the translators in this cluster have more than eight years' experience in translating business intelligence and this represents 87.5 percent of all the translators in the whole group of 24. This percentage is significantly higher than in the global group of translators ($t=3.57$, $p<0.001$).

4. *What tasks does your work involve?*

All translators in this cluster work in post-editing and this represents 53 percent of the total of translators that work in post-editing. This percentage is significantly higher than in the global group of translators ($t=2.38$, $p=0.009$).

6.3.3. Features of Cluster 3

This group is characterized by having experience in translation, but none or less experience in post-editing MT output. The translators in this cluster are: 2, 3, 6, 17, 21 and 22. To these questions, the translators responded as follows:

1. *How long have you been post-editing raw machine translated (MT) output?*

83.5 percent of all translators in this cluster do not have experience in post-editing machine translated output, and this represents 83 percent of the total of translators in the whole group. This percentage is significantly higher than in the global group of translators ($t=3.15$, $p=0.001$).

2. *Please estimate the percentage, on average, that post-editing MT output represents in your work?*

83.5 percent of the translators in this cluster do not post-edit, and this represents 83 percent of the total of translators that do not post-edit in the group of 24. This percentage is significantly higher than in the global group of translators ($t=3.15$, $p=0.001$).

3. *What tasks does your work involve?*

None of the translators in this cluster work in post-editing, which is 66.7 percent of all translators that do not post-edit. This percentage is significantly higher than in the global group of translators ($t=3.23$, $p=0.001$).

6.3.4. Features of Cluster 4

This group of translators is characterized by being young and having less experience. It includes two translators: 10 and 19. To these questions, the translators responded as follows

1. *How long have you been translating business intelligence software?*

Both translators in this cluster have less than two years' experience translating business intelligence and this represents all translators with less than two years' experience in the whole group of 24. This percentage is significantly higher than in the global group of translators ($t=2.69$, $p=0.004$).

2. Age

Both translators in this cluster are less than 25 years old and this represents all the translators that are 25 years old in the whole group of 24. This percentage is significantly higher than in the global group of translators ($t=2.69$, $p=0.004$).

6.4. Experience vs. processing speed

Now that four clusters of translators from the sample of 24 translators have been defined, we can test our third hypothesis, which claims that *the greater the experience of the translator, the greater the productivity in post-editing MT match and Fuzzy match segments*. We can also test the sub-hypothesis that *this experience will not have an impact on the quality (measured in number of errors)*.

6.4.1. Experience vs. processing speed: Fuzzy match

The Fuzzy match values are taken for all translators and their speed is calculated taking the words per minute in Fuzzy match segments for the different clusters:

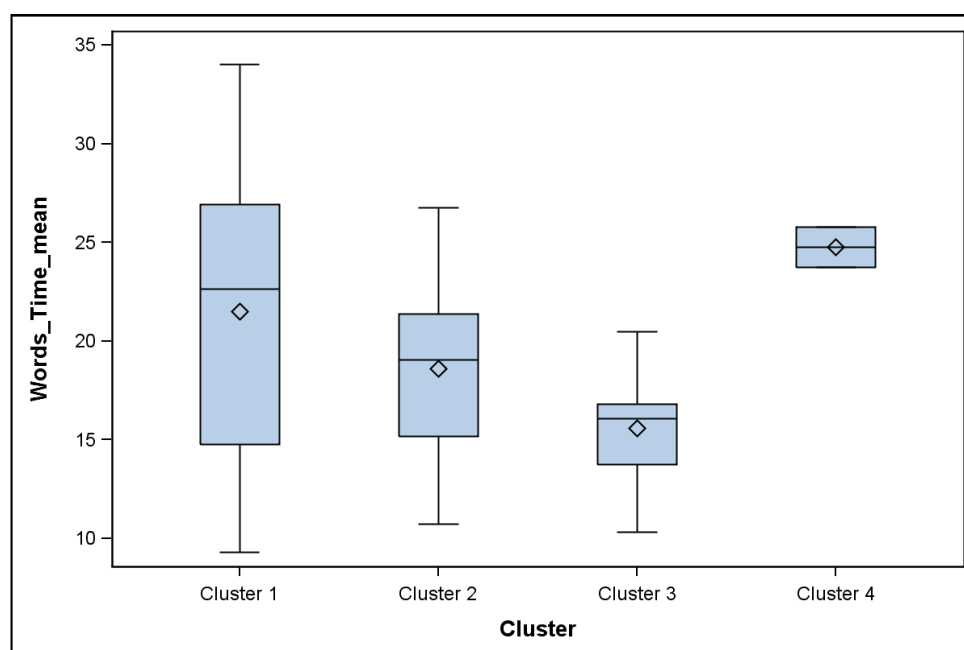


Figure 42: Processing speed vs. Fuzzy match

Cluster 3, the group with no or little experience in post-editing, shows lower processing speed in Fuzzy match (if we look at means and medians) than the other clusters. Cluster 1, the second in overall experience, has a higher mean and median

values than Clusters 2 and 3. Cluster 2, the most experienced, behaves similarly to Cluster 1 but slower than Cluster 4, which has a very homogeneous speed (only two translators) and the highest mean and median values. Cluster 1 shows more deviation, with the slowest values (Translator 11) but also the fastest values (Translator 13 is in this group and she was the fastest translator as described in Chapter 4). Let us look at the descriptive data in Table 57.

Cluster	N	Mean	Median	SD	Min	Max
Cluster 1	8	21.49	22.65	8.41	9.29	34.03
Cluster 2	8	18.59	19.05	4.95	10.73	26.74
Cluster 3	6	15.58	16.07	3.37	10.33	20.48
Cluster 4	2	24.76	24.76	1.43	23.75	25.78

Cluster	N	Min	Q 1	Median	Q 3	Max	Range	Q Range
Cluster 1	8	9.29	14.75	22.65	26.90	34.03	24.75	12.16
Cluster 2	8	10.73	15.18	19.05	21.39	26.74	16.02	6.20
Cluster 3	6	10.33	13.75	16.07	16.82	20.48	10.15	3.07
Cluster 4	2	23.75	23.75	24.76	25.78	25.78	2.03	2.03

Table 57: Processing speed vs. Fuzzy match

Cluster 1 has the second highest mean and median values with the highest deviation, as we saw earlier: Translator 13 has the maximum value at 34.03 and Translator 11, the lowest at 9.29. Cluster 2 has slightly lower figures: Translator 5 has the maximum value 26.74 and Translator 4 the minimum 10.73. Cluster 3 has the lowest values: Translator 3 with 10.33 words and Translator 21 with 20.48 words. Cluster 4 has the highest mean and median values and is the most homogenous group: Translator 19 with 25.78 words and Translator 10 with 23.75 words.

Therefore, if Fuzzy matches are examined in the groups with more experience (Cluster 1 and 2) the productivities are high. However, productivities are also high in Cluster 4, the group with the least experience. The interesting data point in this case is that Cluster 3, with no or little experience in post-editing although with experience on the other areas, has a lower processing speed than the other three clusters. This might indicate that this particular group was slower when translating because their typing speed was slower (the two slowest typists are in this group) or because they invested more time in producing a better translation (we will see this in the following section when we look at the errors per cluster). But how did the clusters then behave with MT matches? Was this Cluster 3, with no experience in post-editing, also the slowest in this category?

6.4.2. Experience vs. processing speed: MT match

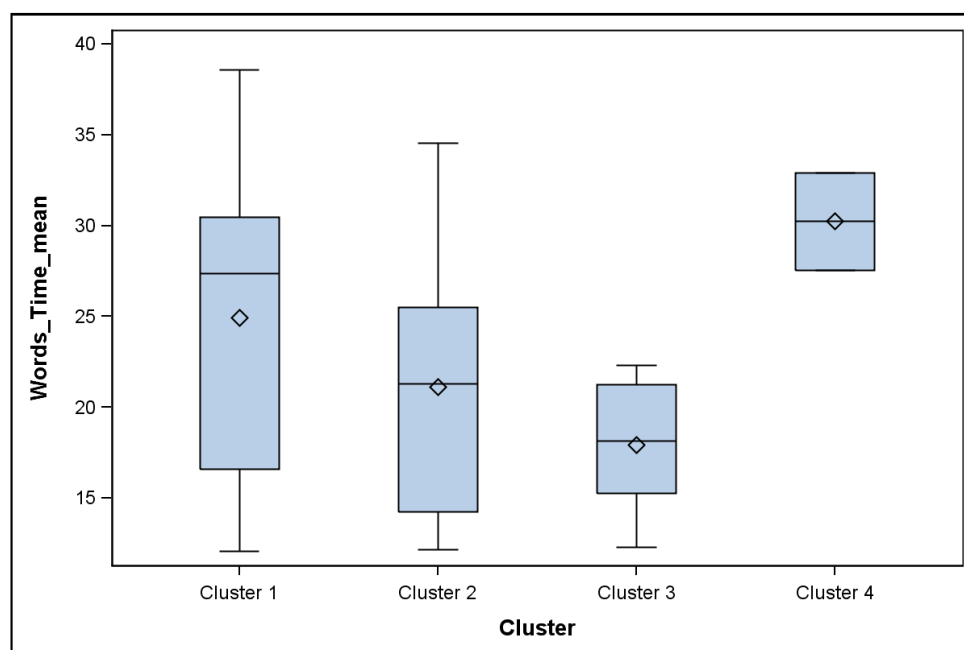


Figure 43: Processing speed vs. MT match

Cluster 4, the group with the least experience, seems to have taken full advantage of MT matches, with very high median and mean values (over 20 words per minute). Cluster 1 and Cluster 2, the groups with most experience, show similar values, although Cluster 1 seems to be slightly faster. There are translators in Clusters 1 and 2 that seem to have quite different speeds, and some show lower speeds. Cluster 3, the group with no post-editing experience, has more homogenous values and again the lowest mean and median values. This is understandable if they declare having no experience in post-editing MT. Let us look at the descriptive data (Table 58) to gain better understanding of the figure above.

Cluster	N	Mean	Median	SD	Min	Max
Cluster 1	8	24.94	27.38	9.09	12.07	38.58
Cluster 2	8	21.10	21.29	7.57	12.17	34.57
Cluster 3	6	17.90	18.14	3.82	12.31	22.33
Cluster 4	2	30.23	30.23	3.79	27.55	32.91

Cluster	N	Min	Q 1	Median	Q 3	Max	Range	Q Range
Cluster 1	8	12.07	16.59	27.38	30.47	38.58	26.51	13.89
Cluster 2	8	12.17	14.26	21.29	25.50	34.57	22.40	11.24
Cluster 3	6	12.31	15.26	18.14	21.24	22.33	10.02	5.97
Cluster 4	2	27.55	27.55	30.23	32.91	32.91	5.36	5.36

Table 58: Processing speed vs. MT match

Cluster 4 clearly has high processing speeds when dealing with MT matches: Translator 19 has the maximum value with 32.91 words per minute, and Translator 10 has the minimum value with 27.55 words, almost identical in the group. Cluster 3, with no post-editing experience, has the lowest values if the mean and median values are considered, but Translator 6 has a speed of 22.33 words per minute, while the rest are all below 22.33 and above 12.31, the range here being smaller than in Clusters 1 and 2. Clusters 1 and 2 have similar minimum and maximum values, although Cluster 1 shows faster mean and median values.

Cluster 4, the group with the least experience, shows the highest mean and median values. This seems to be quite the opposite of what our hypothesis was trying to test. This group is young and has very little experience but they seem to benefit considerably from MT. Nevertheless, we also see that experience seems to be a factor. Cluster 3, the slowest, had no or little post-editing experience. This seems to indicate that younger translators might find it easier to deal with MT post-editing because they might have had more contact with machine translation or translation memory outputs since they started working professionally (we saw, when defining the clusters, that Translators 10 and 19 had the same experience in localization as in post-editing, which shows that they have almost a parallel experience in both areas, while more senior translators do not). At any rate, Clusters 1 and 2, with more experience, still have the highest values at 38.58 (Translator 13) and 34.57 (Translator 5) words per minute respectively. Again, we observe here that post-editing experience is a positive factor if speed is considered. Overall experience can have different influences. On the one hand, translators with more experience can perform well, on the other, translators with less experience can also make good use of MT segments (especially if exposed to or trained in machine translation post-editing).

It will be interesting to see how these four clusters perform when translating on their own, to find out if the different productivities were also related to their own (intrinsic) speed in No match words.

6.4.3. Experience vs. processing speed: No match

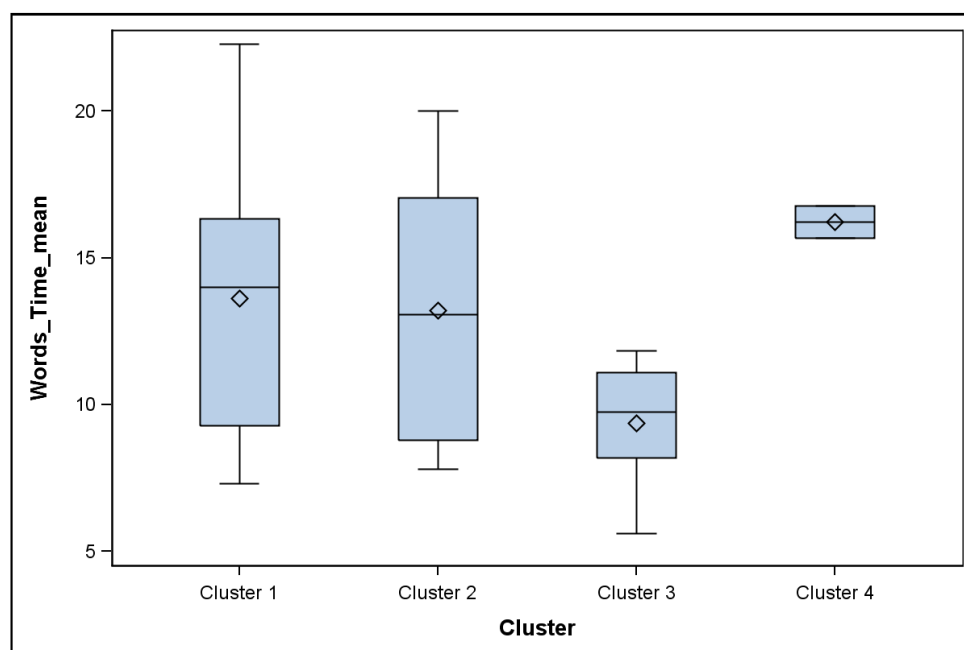


Figure 44: Processing speed vs. No match

Cluster 4 has the highest mean and median values for the No match category. These two translators seem to work at a reasonable speed also when working without a translation aid. Cluster 1 is the second fastest in mean and median values and also seems to have the maximum value in words per minute. Cluster 2 has similar values with a wider range in the quartiles than Cluster 1. Cluster 3 is the group with the lowest mean and median values, and also includes the translator with the lowest value in all the clusters. Table 59 shows the descriptive data.

Cluster	N	Mean	Median	SD	Min	Max
Cluster 1	8	13.61	14.00	5.00	7.32	22.29
Cluster 2	8	13.20	13.08	4.70	7.80	20.00
Cluster 3	6	9.37	9.75	2.24	5.60	11.85
Cluster 4	2	16.24	16.24	0.77	15.69	16.78

Cluster	N	Min	Q 1	Median	Q 3	Max	Range	Q Range
Cluster 1	8	7.32	9.29	14.00	16.34	22.29	14.97	7.06
Cluster 2	8	7.80	8.78	13.08	17.05	20.00	12.20	8.27
Cluster 3	6	5.60	8.18	9.75	11.09	11.85	6.25	2.91
Cluster 4	2	15.69	15.69	16.24	16.78	16.78	1.09	1.09

Table 59: Processing speed vs. No match

In Cluster 4, Translator 19, with 16.78 words per minute, and Translator 10, with 15.69 words per minute, have the highest processing speeds if we look at the median and mean values. However, Translator 13, with 22.29 words per minute, has the

maximum value in Cluster 1, followed by Translator 23, with 20 words per minute, in Cluster 2. The translators in Cluster 3 present lower values overall: Translator 3 has the lowest value at 5.60 and Translator 21 has the maximum value at 11.85 (words per minute), and the range is narrower, meaning that there was more homogeneity in the translators' speeds. All the translators in this cluster were in Speed groups 1 and 2 (as we saw in the Chapter 4).

It seems understandable that Cluster 3 also had low processing speeds when working with MT and Fuzzy matches, since their baseline (No match translation) is within a low speed range. It is, therefore, not clear if their low productivity in the three match categories (Fuzzy, MT and No match) was due to their speed as translators, to lack of experience in post-editing MT output (the lack of familiarity with these types of errors might decrease their speed) or simply because they had spent more time in correcting errors. It is also interesting to note that all the translators that declare having an average typing speed of 0-20 words per minute are in this group (Translators 2 and 22) and the others in the group declared having 41-60 words per minute (Translators 21, 17, 6 and 3).

Table 60 shows the distribution between Clusters and Speed groups (in No match). Group 1 processed less than 10 words per minute; Group 2, from 10 (included) to 15 words per minute; and Group 3, 15 words per minute or more.

Cluster	Group 1		Group 2		Group 3		Total
	N	Row %	N	Row %	N	Row %	
Cluster 1	2	25.00	3	37.50	3	37.50	8
Cluster 2	3	37.50	1	12.50	4	50.00	8
Cluster 3	4	66.67	2	33.33	.	.	6
Cluster 4	2	100.00	2

Table 60: Clusters and Speed groups

Table 60 shows that Cluster 1 contains 25 percent of the slowest translators in No match (Translators 11 and 14); 37.5 percent of medium speed translators (Translators 8, 16 and 18), and 37.5 percent of the fastest translators (Translators 1, 13 and 15). Cluster 2 has 37.5 percent of slowest translators (Translators 4, 9 and 12), 12.5 percent of medium speed translators (Translator 24), and 50 percent of the fastest translators (Translators 5, 7, 20 and 23). Cluster 3, however, has the highest percentage of slowest translators (Translators 2, 3, 6 and 22), 33.33 percent of medium speed translators (Translators 17 and 21) and none of the fastest translators. Cluster 4 has 100 percent of the fastest translators (Translators 10 and 19).

By looking at the descriptive data it is difficult to know if experience made a statistically significant difference in processing speed, although Cluster 3 had the slowest translators in the No match category and it was the cluster that had no experience in post-editing. A linear regression model with repeated measures was applied to the data, taking *logarithm of Words per minute* as the response variable, and *Match category* and *Cluster* as explanatory variables. There are statistically significant differences ($F=169.91$ and $p<0.0001$) between the three translation categories: Fuzzy match, MT match and No match. This is expected, as this was seen when we analyzed productivity. However, there are no statistically significant differences between Clusters, and in the interaction between Clusters and Match category. From this model, mean value estimations were calculated taking the variable *logarithm of Words per minute* according to the Match and Cluster. We present the estimated mean value with their corresponding confidence intervals of 95 percent. The estimations are expressed in words per minute for a better understanding.

Cluster	Mean	Lower	Upper
Cluster 1	18.09	14.27	22.91
Cluster 2	16.46	12.99	20.86
Cluster 3	13.46	10.24	17.69
Cluster 4	22.95	14.30	36.84

Table 61: Estimated mean per Cluster

Although the estimated mean for Cluster 4 is the highest, followed by Clusters 1, 2 and Cluster 3, there are no statistically significant differences between the four clusters. The gap between Cluster 3 (the slowest with 13.46) and Cluster 4 (the fastest with 22.95) is approximately 9 words. The lower and upper intervals overlap with each other, showing that the translators in each cluster presented a variety of speeds not necessarily related to experience. This is contrary to the findings from De Almeida and O'Brien (2010) where faster translators were also the ones with more experience. However, they report on a pilot project with three participants per language, and this number makes it difficult to see the effect experience had on speed. Table 62 shows the estimated mean again, but now showing the Match category and the Productivity gain with respect to No match.

Match	Cluster	Estimated mean	Lower	Upper	Productivity gain
Fuzzy match	Cluster 1	19.85	15.56	25.32	55.23%
Fuzzy match	Cluster 2	17.98	14.09	22.93	44.44%
Fuzzy match	Cluster 3	15.26	11.52	20.21	67.49%
Fuzzy match	Cluster 4	24.74	15.21	40.26	52.49%
MT match	Cluster 1	23.31	18.28	29.74	82.35%
MT match	Cluster 2	19.94	15.63	25.44	60.21%
MT match	Cluster 3	17.54	13.24	23.24	92.57%
MT match	Cluster 4	30.11	18.51	49.00	85.59%
No match	Cluster 1	12.79	10.02	16.31	.
No match	Cluster 2	12.45	9.76	15.88	.
No match	Cluster 3	9.11	6.88	12.07	.
No match	Cluster 4	16.23	9.97	26.40	.

Table 62: Estimated mean according to Match and Cluster

Cluster 4 presents the fastest estimated value, followed by Clusters 1, then 2 and finally Cluster 3. Cluster 3 seems to be the group that obtains the most productivity gain with Fuzzy and MT match (with 67.49 percent and 92.57 percent respectively). In this case, the slowest group (grouped according to their experience) does take more advantage of the translation aids as we were trying to see in section 4.4. Cluster 4 obtains more productivity gain with MT (85.59 percent) as opposed to Fuzzy (52.49 percent). Cluster 1 shows more productivity gain with MT (82.35 percent) as opposed to Fuzzy match (55.23 percent). Finally Cluster 2 also shows more productivity with MT (60.21 percent) than with Fuzzy (44.44 percent). Figure 45 clearly illustrates the findings.

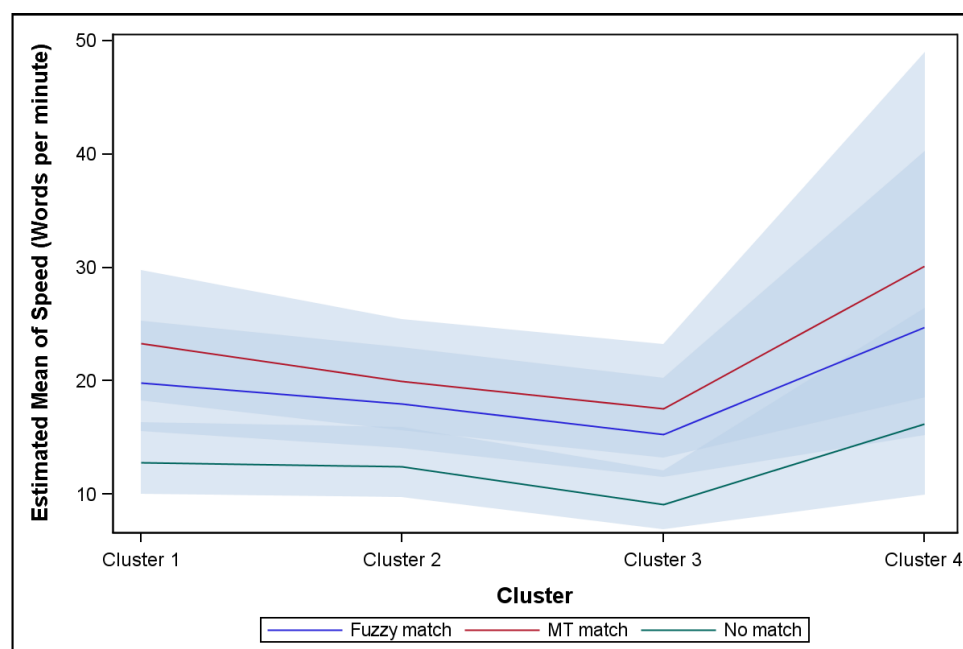


Figure 45: Estimated mean of speed per Cluster and Match

Speed is always lower for Cluster 3, higher for Cluster 4, and similar for Clusters 1 and 2 in the three match categories. No match is significantly different for all clusters (green line), while Fuzzy match (blue line) and MT match (red line) show similar values, except with Cluster 4, where the MT match is slightly higher. To double-test the validity of the findings, non-parametric comparisons were set-up (Kruskal-Wallis analysis of variance) and we found no statistically significant differences between the Clusters according to the Match category if speed was considered.

Consequently, the first part of our hypothesis that says that the greater the experience of the translator, the greater the productivity in post-editing MT match and Fuzzy match segments is not supported in our experiment. Although Clusters 1 and 2, with more experience, show high values, Cluster 4, with less experience, also shows the highest mean and median results. Cluster 3, on the other hand, with no post-editing experience, shows lower speed values, but this was also the case in the No match category. Hence the reason could lie more in their own average typing speed or general processing speed than in the fact that they have no experience in post-editing MT matches.

In the same way that productivity was linked to quality, experience needs to be related to productivity and to quality. Does Cluster 4 present more errors than Cluster 3, for example?

6.5. Experience vs. number of errors

In this section, we will look at the error results according to the different clusters. As we did with processing speed, Fuzzy matches will be examined first.

6.5.1. Experience vs. number of errors: Fuzzy matches

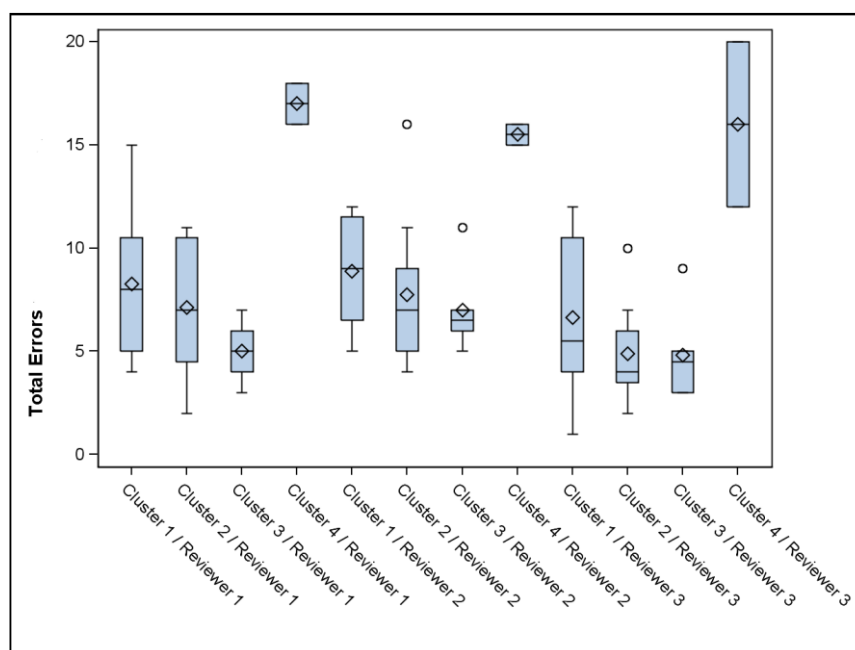


Figure 46: Total errors for Fuzzy match in clusters

Interestingly, Cluster 4 has the highest number of errors according to all three reviewers, indicating that this Cluster was the fastest if the mean value is considered, but it was not as rigorous or thorough when editing the Fuzzy match category. On the other hand, Cluster 3 has the lowest number of errors, indicating that this Cluster was the slowest but also thorough when processing the Fuzzy match segments. The differences between Clusters 1 and 2 are not pronounced. Reviewer 1 places Cluster 4 first in number of errors, and the other three clusters are not that dissimilar. The cluster with the lowest value is Cluster 2 (Translator 20 with 2 errors), and Cluster 1 has quite a high value (Translator 18 with 15 errors). Reviewer 2 also places Cluster 4 first in number of errors, and the other three clusters have similar values. There are also some outliers here: Translator 7 with 16 errors is in Cluster 2 (at the same level as Cluster 4) and Translator 3 with 7 errors is in Cluster 3. Reviewer 3 also places Cluster 4 at the top of errors, reaching 20 errors. The other three clusters have similar values: Cluster 1 again has the lowest value (Translator 15 only has one error). Table 63 shows the descriptive data for the number of errors in Fuzzy matches.

Cluster	Reviewer	N	Mean	Median	SD	Min	Max
Cluster 1	Reviewer 1	8	8.25	8.00	3.77	4.00	15.00
	Reviewer 2	8	8.88	9.00	2.90	5.00	12.00
	Reviewer 3	8	6.63	5.50	3.96	1.00	12.00
Cluster 2	Reviewer 1	8	7.13	7.00	3.48	2.00	11.00

Cluster	Reviewer	N	Mean	Median	SD	Min	Max
Cluster 3	Reviewer 2	8	7.75	7.00	3.99	4.00	16.00
	Reviewer 3	8	4.88	4.00	2.53	2.00	10.00
	Reviewer 1	6	5.00	5.00	1.41	3.00	7.00
Cluster 4	Reviewer 2	6	7.00	6.50	2.10	5.00	11.00
	Reviewer 3	6	4.83	4.50	2.23	3.00	9.00
	Reviewer 1	2	17.00	17.00	1.41	16.00	18.00
	Reviewer 2	2	15.50	15.50	0.71	15.00	16.00
	Reviewer 3	2	16.00	16.00	5.66	12.00	20.00

Table 63: Total errors for Fuzzy match in clusters

Cluster 4 has the highest mean values for all three reviewers, the highest median values, and the highest minimum and maximum values. The only similar maximum value is in Cluster 2: Translator 7 with 16 errors according to Reviewer 2. Cluster 3 has the lowest mean and median values from the three reviewers. However, the minimum and maximum values are very similar in these three clusters (1, 2 and 3), indicating that some translators had low or high values irrespective of the cluster they were in. When the Global error database is consulted, Translator 10 and Translator 19 (Cluster 4) made more mistakes in Terminology. Translator 10 has 26 Terminology errors and Translator 19, 31 (aggregated value from all three reviewers), and 57 Terminology errors in the whole Cluster. This clearly indicates that translators in Cluster 4 gained speed because they tended not to check the glossary. They accepted the terminology as it was presented to them in the Fuzzy matches. They also have 16 Accuracy errors and 14 Language errors. On the other hand, Cluster 3, with six translators instead of two (as in Cluster 4), has 28 Terminology errors and 43 Language errors, indicating that the translators were more thorough when checking terminology. Cluster 2, with eight translators, has 47 Terminology errors (the same number of errors as Cluster 4 but with four times the number of translators) followed by Accuracy with 44 errors, and Language 40. Cluster 1 with eight translators has 75 Terminology errors, followed by Accuracy, 47 errors and Language 45 errors. We observe that Cluster 3 was slowest because they might have devoted more time to check the terminology against the glossary provided.

For Fuzzy matches, the results are rather clear. Cluster 4, with less experience and higher speed, left or made more errors in the segments according to the three reviewers. Cluster 3 made slightly less, although results for Clusters 1, 2 and 3 are quite similar. These results are interesting since they seem to signal a lack of attention to certain important aspects of the translation process in the more novice translators. We

suspect that this would be the case for the whole assignment, but let us have a look at the results for the MT matches in Figure 47.

6.5.2. Experience vs. number of errors: MT matches

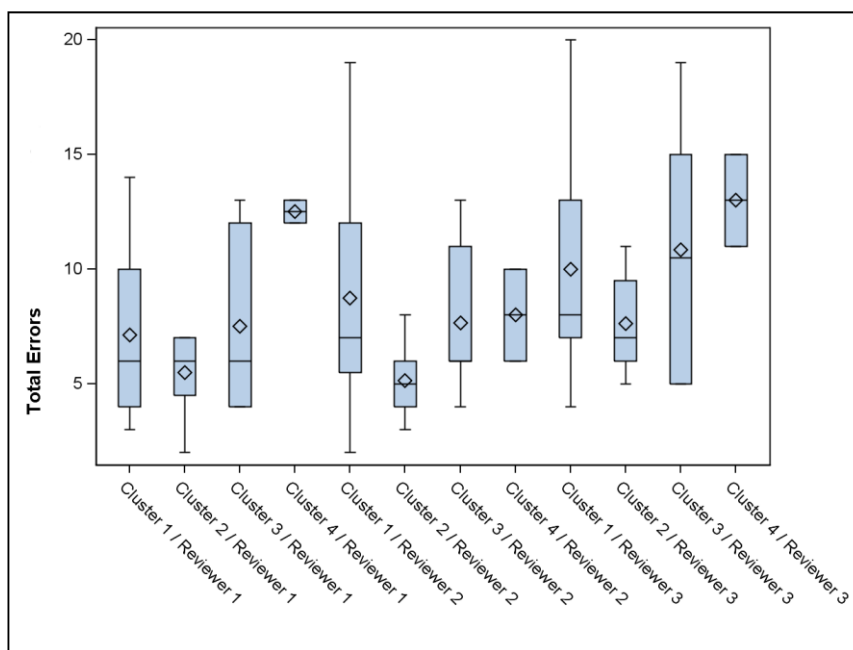


Figure 47: Total errors for MT match in clusters

These results are particularly interesting. In this case, the differences between the clusters are not as pronounced as with the Fuzzy matches. We think this is possible because, as we saw in the TER section, some of the MT matches were perfect matches, with no changes required, and although translators can still introduce mistakes, it would be logical that if the translators in Cluster 4 had problems in terminology (failing to check the glossary consistently, and a certain lack of understanding of instructions), the perfect matches could help them lower the number of errors. Reviewer 1 places Cluster 4 in the top range of errors, but Translator 13 in Cluster 1 with 14 errors and Translator 3 in Cluster 3 (coincidentally this translator has less experience in localization) with 13 errors are very close to the values in Cluster 4. Cluster 2, the most experienced group, seems to perform well with MT matches. Translator 9, with 2 errors, has the lowest value in this category. Reviewer 2, on the other hand, has all clusters with similar number of errors. Cluster 1 has the extreme values (Translator 13 with 19 errors). Cluster 2 has lower values because Translator 9 and Translator 23 have only 3 errors. Reviewer 3 has Cluster 4 and Cluster 3 at almost the same level. Cluster 1 has the highest values (Translator 13 with 20 errors) but also Translator 15 has the lowest value

(5 errors) in this same cluster. Cluster 3 does not perform as well as in the Fuzzy match category (perhaps because they had no or little experience in post-editing). Translator 22 has 19 errors in this category, unlike other translators (Translators 2 and 17 have 5 errors) in this same cluster. Table 64 shows the descriptive data for MT match.

Cluster	Reviewer	N	Mean	Median	SD	Min	Max
Cluster 1	Reviewer 1	8	7.13	6.00	4.19	3.00	14.00
	Reviewer 2	8	8.75	7.00	5.60	2.00	19.00
	Reviewer 3	8	10.00	8.00	5.21	4.00	20.00
Cluster 2	Reviewer 1	8	5.50	6.00	1.77	2.00	7.00
	Reviewer 2	8	5.13	5.00	1.64	3.00	8.00
	Reviewer 3	8	7.63	7.00	2.13	5.00	11.00
Cluster 3	Reviewer 1	6	7.50	6.00	4.04	4.00	13.00
	Reviewer 2	6	7.67	6.00	3.50	4.00	13.00
	Reviewer 3	6	10.83	10.50	5.60	5.00	19.00
Cluster 4	Reviewer 1	2	12.50	12.50	0.71	12.00	13.00
	Reviewer 2	2	8.00	8.00	2.83	6.00	10.00
	Reviewer 3	2	13.00	13.00	2.83	11.00	15.00

Table 64: Total errors for MT match in clusters

Cluster 2 has the lowest mean values and Cluster 4 the highest if we consider all three reviewers. However, not all the values are as different as what we saw in the Fuzzy match category. Cluster 4 has the highest minimum values: there are only two translators in this cluster and they behave similarly, but the maximum values are to be found in Cluster 1 (Translator 13 with the highest values). If we look at the Global error database to see the type of errors each Cluster made the results are different from those found in Fuzzy matches. There are Terminology errors but here the majority of errors are on Language overall, according to all three reviewers. The reviewers seem to be of the opinion that not enough changes were made in the segments for them to be linguistically acceptable. Cluster 1, with eight translators, has 59 Language errors, 39 Terminology errors, 40 Mistranslation, 41 Accuracy. Cluster 2 has the highest number of errors in Language with 44 errors, Terminology and Accuracy with 20, Mistranslation with 30. Cluster 3, with six translators, has 66 errors in Language, 19 in Terminology and 18 in Accuracy, 25 in Mistranslation. Cluster 4, with only two translators, has 21 in Language and 28 in Terminology, 4 in Mistranslation and 3 in Accuracy. This seems to indicate that the least experienced translators still did not check the glossary with MT matches: the number of errors might be lower simply because the proposals were correct, perfect matches. Cluster 2, the most experienced group, performed better with MT matches with fewer errors and fewer Language errors than the other groups. Hence, this might indicate that experience is a factor when dealing with MT matches, but also that the differences in errors between the clusters were not as

pronounced as in Fuzzy matches. Cluster 4 performed faster with MT matches and the number of errors was lower than with Fuzzy matches, and this might indicate that with translators who have less experience, high quality output MT might be a better option than translation memories below the 94 percent threshold.

If translators behave differently with Fuzzy than with MT matches, how did they do without any translation proposal? Figure 48 shows the results for the No match category.

6.5.3. Experience vs. number of errors: No matches

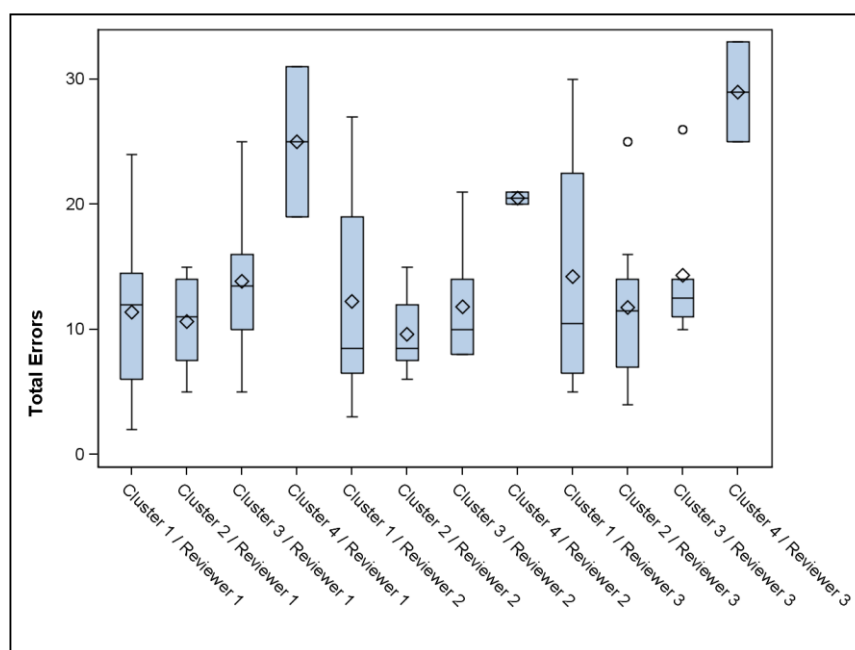


Figure 48: Total errors for No match in clusters

The results here are more similar to the Fuzzy match than to the MT match results. Cluster 4 clearly has the highest number of errors, and the other three clusters are very close in results. Once again, Cluster 2 seems to have the most homogenous data, thus indicating that this group did not have translators with extreme values as in Clusters 1 and 3. Reviewer 1 found many errors in Cluster 4, more for Translator 10 (31 errors) than for Translator 19 (19 errors), therefore we can see a wider block. In Cluster 1 we observe translators with few errors (Translator 11 with 2 errors for example) while others have more (Translator 18 has 24 errors). Cluster 3 also has a very high value, (Translator 3 with 25 errors), and at the other end Translator 6 with only 5 errors. Reviewer 2 found more errors in Cluster 4 overall but not as many as the other two reviewers. Cluster 1 has the highest value here, as Translator 13 has 27 errors, followed

by Translator 18 with 20. Once more, Cluster 3 presents a high value because Translator 3 has 21 errors. For Reviewer 3, however, Cluster 4 has high values: 33 errors for Translator 10, and 25 for Translator 19. However, Cluster 1 with Translator 13 (30 errors) and Translator 18 (25 errors) is not very far behind. Cluster 3, on the other hand, performs well except for Translator 3 with 26 errors, which pushes up the mean value. Cluster 2 has the lowest values (Translator 12 with 4 errors), but Translator 4 with 25 is an outlier. Table 65 shows the descriptive values for the No match category.

Cluster	Reviewer	N	Mean	Median	SD	Min	Max
Cluster 1	Reviewer 1	8	11.38	12.00	6.86	2.00	24.00
	Reviewer 2	8	12.25	8.50	8.38	3.00	27.00
	Reviewer 3	8	14.25	10.50	9.68	5.00	30.00
Cluster 2	Reviewer 1	8	10.63	11.00	3.93	5.00	15.00
	Reviewer 2	8	9.63	8.50	3.11	6.00	15.00
	Reviewer 3	8	11.75	11.50	6.56	4.00	25.00
Cluster 3	Reviewer 1	6	13.83	13.50	6.68	5.00	25.00
	Reviewer 2	6	11.83	10.00	5.04	8.00	21.00
	Reviewer 3	6	14.33	12.50	5.89	10.00	26.00
Cluster 4	Reviewer 1	2	25.00	25.00	8.49	19.00	31.00
	Reviewer 2	2	20.50	20.50	0.71	20.00	21.00
	Reviewer 3	2	29.00	29.00	5.66	25.00	33.00

Table 65: Total errors for No match in clusters

Cluster 4 clearly has the highest mean and median values according to all reviewers. They also have a very high minimum value. Reviewer 2 found an almost identical number of errors for both translators. Cluster 3 has higher aggregated values, but all three clusters have similar median and mean values, showing that many translators have similar numbers of errors. If we look at the Global error database to see the type of errors each Cluster made, the results are slightly different from those found for Fuzzy and MT matches. The majority of errors are in Language, followed by Terminology and Style. The reviewers seem to be of the opinion that the segments were not linguistically acceptable, as with MT matches. However, when we look at Cluster 4, the majority of errors are in Terminology (54 errors). Once again, the glossary and the instructions were not followed correctly. It also has 43 Language errors and 24 Style errors. Accuracy and Mistranslations rank lower with 15 and 13 respectively. Cluster 3 has 122 Language errors (but Translator 3 alone has more than 40), 43 Terminology errors (with 6 translators as opposed to 2 as in Cluster 4), 40 Style errors, 19 Accuracy errors, and 12 Mistranslation errors. In this case, the problem is not so much in terminology as in Language. Clusters 1 and 2 have similar values. Language errors are the highest (106 and 93 respectively) followed by Terminology, Style, Accuracy and Mistranslation.

The number of errors in Clusters 1, 2 and 3 are similar. This makes sense since these groups of translators are very experienced overall. Notwithstanding this, there are some with more experience in certain areas. Cluster 4, the one with the least experienced translators, gave a poorer performance in this category, indicating that experience when translating without any translation aid influences the number of errors. This seems to point to the fact that translators with experience work better with the instructions given and are more thorough. This was also true for Fuzzy matches and to a lesser extent for MT matches.

Are these differences significant? We saw differences in speed but these were not statistically significant between the Clusters, so what will be the case for the number of errors? A Poisson regression model is applied with repeated measures taking the variable *Total errors* as the response variable and the offset as *text length*. Statistically significant differences are observed for the variable *Total errors* between the different Match categories: Fuzzy, MT and No match ($F=53.50$ and $p<0.0001$), as well as for the different clusters ($F=7.61$ and $p<0.0001$). Finally, statistically significant differences are observed in the interaction between Match categories and Clusters ($F=3.37$ and $p=0.0039$).

From this model, estimations of the mean values are obtained for the variable (total errors /text length) according to Match category with the corresponding interval levels of 95 percent. We present the results of these estimations but expressed in number of errors per segment length for better understanding. We consider the length of the original text (Fuzzy match, 618 words, MT match, 757 words and No match 749 words).

Match category	Mean	SD	Lower	Upper
Fuzzy match	8.02	0.51	7.06	9.10
MT match	8.16	0.53	7.18	9.27
No match	14.05	0.80	12.55	15.73

Table 66: Estimated mean of errors per Match categories in clusters

The estimated mean of errors in the original text for Fuzzy match is 8.02 and the confidence interval is (7.06 and 9.10). For MT match it is 8.16 and the confidence interval is (7.18 and 9.27). Finally, for No match, the estimated mean is 14.05 and the confidence interval is (12.55 and 15.73). In fact, this is what we have seen above when we were analyzing the quality of the translations. There are no statistically significant

differences in the number of errors between MT and Fuzzy match but there are between No match and the other two categories.

Cluster	Mean	SD	Lower	Upper
Cluster 1	26.82	2.12	22.94	31.36
Cluster 2	22.10	1.79	18.83	25.95
Cluster 3	25.17	2.32	20.97	30.21
Cluster 4	49.30	7.31	36.77	66.11

Table 67: Estimated mean of errors in cluster

The estimated mean of errors in Cluster 1 is 26.82 and the confidence interval is (22.94, 31.36). In Cluster 2 the mean is 22.10 and the confidence interval (18.83, 25.95). In Cluster 3 the mean is 25.17 and the confidence interval is (20.97, 30.21). Finally, the estimated mean for Cluster 4 is 49.30 and the confidence interval is (36.77, 66.11). This result is interesting and different from what we saw in speed (see section 6.4). This is in line with the findings from De Almeida and O'Brien (2010) where more experienced translators were more accurate. Here Cluster 4 shows a statistically significant difference in the number of errors with respect to the other three clusters.

Match	Cluster	Mean	SD	Lower	Upper
Fuzzy match	Cluster 1	7.41	0.74	6.08	9.03
Fuzzy match	Cluster 2	6.41	0.67	5.21	7.89
Fuzzy match	Cluster 3	5.42	0.69	4.21	6.96
Fuzzy match	Cluster 4	16.04	2.72	11.47	22.44
MT match	Cluster 1	8.07	0.79	6.65	9.79
MT match	Cluster 2	5.93	0.64	4.79	7.33
MT match	Cluster 3	8.37	0.94	6.70	10.45
MT match	Cluster 4	11.08	2.02	7.72	15.90
No match	Cluster 1	11.81	1.06	9.89	14.10
No match	Cluster 2	10.39	0.96	8.65	12.48
No match	Cluster 3	12.87	1.31	10.52	15.75
No match	Cluster 4	24.65	3.91	18.01	33.73

Table 68: Estimated mean of errors per match and cluster

When we observe the interaction between Clusters and Match categories in Table 68, the results are interesting once again. Cluster 4 shows statistically significant differences in the Fuzzy match and No match categories. But in the MT match category, although the number of errors is higher, the confidence intervals overlap (row 8), showing that this difference is not statistically significant in this particular match category. So MT, in this instance, acted as a “leveler” in terms of errors for Cluster 4. We can see this clearly in Figure 49.

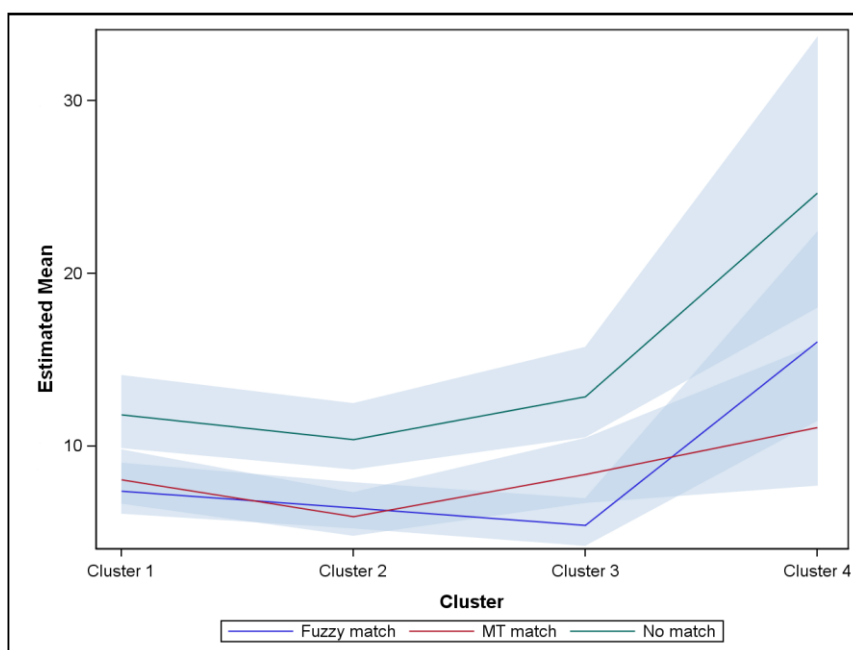


Figure 49: Estimated mean of errors and clusters

For each match category (Fuzzy, MT and No match), Cluster 4 has a higher estimated mean of errors than do the other clusters. Further, this estimation shows statistically significant differences with respect to the other clusters in the case of Fuzzy match and No match. For the MT match category, the estimated mean is higher in Cluster 4 but this estimation is not statistically different from the other clusters. We observe how the green line (No match) is the highest (more errors), it bends slightly down at Cluster 2 and then continues to rise in Cluster 3 and jumps up in Cluster 4. The red line (MT match) is lower for Cluster 2, but slightly higher for Cluster 1, 3 and 4. The blue line (Fuzzy match) is lower for Cluster 1 and particularly Cluster 3 (the ones with no or little experience in post-editing) and jumps up again in Cluster 4.

The second part of our hypothesis claims that *experience will not have an impact on the quality (measured in number of errors)*. Now, after going through the results, we find that this hypothesis is not supported by our data. In fact, the results show the opposite, that experience does play a part in the number of errors found. It is true that for Clusters 1, 2 and 3 there are no statistically significant differences, but then again these three groups have very similar experience, while Cluster 4 was decidedly less experience. This group made more mistakes, mainly because they did not follow instructions and hence avoided the glossary, resulting in a higher speed but poorer quality. Interestingly, the number of errors was not as high in MT match segments, and this could be (as we saw in the TER section) because some segments in MT required

little change or because the terminology was already consistent with the glossary. Cluster 2, the most experienced group, has fewer errors although these were not significantly lower. Cluster 3, with no experience in post-editing, performed worse in this category, showing again that training and experience in this task helps not only with respect to speed but also in quality.

6.6. Conclusions on the translators' experience

This group of professional translators is quite homogenous in terms of experience, with slightly different areas of expertise. They also have considerable experience at using tools and some experience in post-editing MT output, although the task represents a low percentage of their work and has not been performed for a very long period of time. Their working speed seems to be in accordance with industry standards and is quite homogeneous. A multivariate analysis was setup to distribute the translators into four different clusters to test our hypothesis. The results indicate that the incidence of experience on the processing speed is not significantly different for this group of translators. Translators with more experience performed similarly to other very novice translators. Translators with less or no experience in post-editing were the slowest group but again the differences were not significant. This seems to be different from our previous findings (Guerberof 2008) and from the findings by De Almeida and O'Brien (2010), although more in line with the findings in Tatsumi (2010). However, the numbers of participants in those studies are lower, to the extent that one post-editor has a great impact in the whole group, whereas in this project there were 24 translators with different experience and also speed. Further research is needed to draw definitive conclusions.

Our findings on errors are in line with those in De Almeida and O'Brien (2010). Translators with more experience made fewer mistakes than those with less experience. As Offersgaard et al. (2008) suggests a "good post-editor is an experienced proof-reader" (ibid: 156). The number of errors was significantly different between Cluster 4 (the translators with the least experience) and the other clusters with regards to Fuzzy and No match. The difference was higher but not significant for MT match. Also the type of errors made by the novice group were mostly Terminology errors, as opposed to Language or Style as in the other clusters, indicating that translators with less experience were less thorough with terminology and with instructions than were the

more experienced groups. But this is not to say that they did not have more errors in the other categories as well. The MT output, however, seems to have a leveling effect as far as errors is concerned. This might lead us to suggest that using high-quality MT output as opposed to Fuzzy matches below the 95 percent threshold might be advisable for a group of translators with less experience, as there are more probabilities of having perfect matches in the proposed texts and hence of making fewer mistakes. Are novice translators more tolerant to errors in quality than senior translators? Our reviewers were senior translators and they might have a different idea of quality than the novice translators. Is the current review method adequate to establish a quality suitable for the market? Lagoudaki (2008) and Flourney and Duran (2009) also suggest that inexperienced translators seem to be more tolerant of MT errors and structures than experienced ones. It might be that “new” generations of translators might have a different outlook on translation quality to that of senior translators. Finally, it was also observed that the cluster with the least or no experience in post-editing performs better with Fuzzy matches in terms of errors than with MT matches, and this seems to indicate that experience and training have a definite pay-off in terms of quality, although this might not be the only factor.

Our third hypothesis was not supported because experience did not have a significant impact on the translators’ speed, although it did on the translators’ quality. This was quite the opposite of what we had seen in our previous study. We feel that results can vary depending on how experience is defined and statistical calculations are made. However, drawing from our working experience, it is often seen that experience, in terms of years of experience and even exposure to certain tasks, is not a guarantee of speed or of quality.

PART III: Qualitative results

This part presents the results from the post-assignment questionnaire that translators and reviewers filled in as well as the data gathered during the debriefings carried out after all other tasks were completed.

Chapter 7: Questionnaire results

This chapter includes the feedback the translators and reviewers gave after completion of the questionnaire. As was explained in section 6.1, the first part of the questionnaire was related to the translators' experience. We have used this information to group the translators into clusters and test if the speed and number of errors were related to their experience. The second and third part of the same questionnaire was related to their opinions on several topics and their opinions on the assignment. We will present the results obtained from these two parts in the following sections.

7.1. Translators' opinions

There were 11 questions in section "About your opinions". We have organized each question into a table containing the responses from the 24 translators. All quotations directly from the translators are in English and they are exactly the same as the participant put them in the on-line questionnaire.

Please estimate how often the following statements describe your revision procedure (you will need to select an option in each statement):						
Answer Options	Never	Rarely	Sometimes	Frequently	Always	Response Count
<i>As I translate, I recheck my translation before going to the next segment.</i>	1	1	2	5	15	24
<i>Immediately after I finish the translation of one file, I go back and review all my translations.</i>	1	4	7	7	5	24
<i>After I finish the translation of all files assigned to me, I review the whole batch of files.</i>	2	1	4	8	9	24
<i>After I finish one day of work, I go back and review all work done in that day.</i>	6	5	7	4	2	24
<i>Other (please specify)</i>						3

Table 69: Question 1: Revision procedures

The data show that the revision procedure this group of translators tends to use most frequently is checking their translation before proceeding to the next segment (as

they had to do with this assignment) and reviewing the whole batch of files assigned after finishing. This is in line with conclusions on revision styles drawn by Dimitrova (2005): “a segment is often revised before going on to the next segment” (ibid: 144). Reviewing at the end of one day’s work seems to be much less frequent. Translator 13 marked “Never” for the first three first options of this question. This might explain why this particular translator had so many errors in the assignment where she was asked to review each segment after completion.

If you have post-editing experience, which of the options below best represents your experience?		
Answer Options	Response Percent	Response Count
<i>My productivity when post-editing has been constant over time.</i>	45.0%	9
<i>My productivity when post-editing has increased over time.</i>	40.0%	8
<i>My productivity when post-editing has decreased over time.</i>	0.0%	0
<i>I do not know.</i>	15.0%	3

Table 70: Question 2: Post-editing learning curve

We were interested in knowing how translators perceived their productivity in post-editing with growing experience. It seems that some of them might perceive an increase in productivity (40 percent) but others do not (45 percent). Two out of the three “I do not know” responses belong to translators that had declared not having experience in post-editing, so this is in keeping with that. None of the translators think that their productivity decreases over time. This response is interesting. Since post-editing can be a very repetitive task (correcting same type of errors over time) and it could be logical that translators perceive a constant or increased productivity. However, precisely because it is very repetitive, this could cause tiredness and potentially result in a decrease in productivity.

If you have post-editing experience, which of the options below best describes your experience?		
Answer Options	Response Percent	Response Count
<i>Experience has not affected my ability to spot MT errors - I correct them the same way as when I started.</i>	30.0%	6
<i>As I acquire more experience it is more difficult for me to detect MT errors, as I have become used to them.</i>	0.0%	0
<i>As I acquire more experience it is easier for me to detect MT errors, as I look for the same patterns.</i>	55.0%	11
<i>I do not know.</i>	15.0%	3

Table 71: Question 3: Post-editing proficiency

Most of the translators (55 percent) think that experience helps them to detect errors when post-editing. Note that 30 percent declared that they correct errors in the same way as when they started. None of the translators think that experience with post-

editing might affect their ability to detect errors. This is quite a positive response from the translators, since it is often suggested by translators that “growing accustomed” to the errors could result in a decline in the overall quality.

If you have post-editing experience, which of the options below best represents your experience? Reviewing means here to go over a human translation, identify and correct errors. Post-editing means here to go over MT out-put, identify and correct errors.		
Answer Options	Response Percent	Response Count
<i>Post-editing, for me, requires the same effort as reviewing human translations.</i>	30.0%	6
<i>Post-editing, for me, requires more effort than reviewing human translations.</i>	40.0%	8
<i>Post-editing, for me, requires less effort than reviewing human translations.</i>	20.0%	4
<i>I do not know.</i>	10.0%	2

Table 72: Question 4: Post-editing effort

Interestingly four translators responded that post-editing required less effort than reviewing human translation (20 percent). These were Translators 8, 9, 10 and 11. Translators 8, 10 and 11 in this project gained higher productivity with MT segments than with Fuzzy matches but Translator 9 did not (this translator skipped certain Fuzzy match segments and this would be a 0 value in seconds, thus increasing her speed in this particular category), as seen in Table 9 (section 4.3). Still, 40 percent (Translators 1, 4, 5, 7, 12, 13 and 20) think there is more effort required in post-editing. Except for Translators 1 and 12, who in this project had higher productivity when editing Fuzzy matches, the participants showed higher productivity with MT matches. It is important to note that effort not only refers to productivity, or time gained - it also implies *cognitive* effort (Krings 2001, O’Brien 2006b). Translators might perceive a higher (cognitive) effort when post-editing and still be more productive than when editing human translations. Furthermore, the data that we obtained in this study do not represent all the experience these translators have in post-editing or reviewing, and as we have pointed out repeatedly in this study, the MT output quality was high in this particular project.

How satisfied are you with the price per word you receive from your customers?							
Answer Options	Highly unsatisfied	Unsatisfied	Satisfied	Very satisfied	Highly satisfied	Rating Average	Response Count
<i>I am</i>	2	7	11	4	0	2.71	24
<i>Comments</i>							6

Table 73: Question 5: Price satisfaction

The average rating for this question is 2.71 out of 5. Although this is above the median 2.5, there are still nine translators that were either “Unsatisfied” or “Highly unsatisfied” (Translators 2, 3, 9, 11, 12, 17, 19, 20 and 23). The comments were varied. Translator 9 mentioned that due to the current economic recession the prices had gone down and that she makes less money now than when she started 15 years ago. Translator 17 commented that it depended on the task or the agency (some offer better prices than others, and this could depend, in turn, on their direct customers). Translator 20 mentioned that MT matches were poorly paid and that in her opinion it took less time to edit fuzzy matches. In this project, Translator 20 performed faster in MT but the difference with Fuzzy matches was not as high as with other translators (20.25 in Fuzzy match as opposed to 23.03 words per minute in MT match). Moreover, this is only one small project and Translator 20 has between four and six years experience in post-editing. As we saw in Chapter 4:, different translators have perform differently in very similar situations: one can be faster when post-editing than when reviewing and therefore a particular payment method might be better suited to that translator than to another. Pricing, however, does appear to be a problematic aspect in the view of this group of 24 translators.

How satisfied are you with the work that you do as a translator (not considering price now)?							
Answer Options	Highly unsatisfied	Unsatisfied	Satisfied	Very satisfied	Highly satisfied	Rating Average	Response Count
<i>I am</i>	0	1	4	13	6	4.00	24
<i>Comments</i>							2

Table 74: Question 6: Job satisfaction

The translators give a 4 average rating for their satisfaction if price is not considered, and 19 are very satisfied or highly satisfied. So price is definitely a factor that causes some dissatisfaction at least in this particular group of translators. Only Translator 9 is clearly unsatisfied with the work done as a translator:

On the whole, I'm working on very tight schedules, with a bad organization on behalf of many customers, in [sic] very short projects (or parts of them) that need way too long to get ready before the actual translation, and I'm receiving no recognition for the good jobs done. Translation is no longer enjoyable to me.

This seems to be a good summary of reasons for being unsatisfied, and it suggests ways to improve the situation. These comments might be shelved as facts translators need to cope with if they work in localization (“this is the way things are”), but we do

believe a lot can be done on this front to make translators' work more satisfying (for example, in the quantity and quality of instructions). Of course, this is a small sample and a comment from only one translator, but the comments seem quite relevant: stress, too many tasks for too little compensation, and anonymity are problems to be dealt with. The other comment came from Translator 5, who enjoyed the challenges the job had to offer. Variety was an important point for this particular translator (variety is a sought after characteristic among translators as remarked in Lagoudaki 2008), who was "Highly Satisfied", and this might address the question of whether post-editing is an activity that can be done continuously throughout eight hours of work. It seems that alternated and new tasks might be a good strategy to keep translators interested and motivated.

How adequate is the standard payment of fuzzy matches in Translation Memories in relation to the productivity you obtain with them? Standard means here that you receive approximately 20 to 30% for 95-99% fuzzy match, repetitions and 100% matches (if required); 60-66% of the word rate for reviewing 75-94% matches; and full rate for 0-74% matches.

Answer Options	Highly unfair	Unfair	Adequate	Advantageous	Highly lucrative	Rating Average	Response Count
<i>I think it is</i>	1	5	17	1	0	2.75	24
<i>Comments</i>							8

Table 75: Question 7: TM pricing

The rating average is 2.75 out of 5, similar to the rating for pricing. This is logical, since currently most localization projects involve translation memories, thus Fuzzy-match payment. A total of 17 translators found the payment adequate, although there were numerous comments stating that it really depended on the quality of the translation memory, the language combination, the text, and the type of project. Translator 17 mentioned that "I rarely get paid for 99% fuzzy matches, repetitions and 100% matches". Translator 13 thought that the payment was "Highly unfair" because:

I pay for the tools to obtain productivity, pay to learn and certify myself on their use, pay to gain experience in the field and pay for faster computers, not the clients, the benefit should be all mine.

This is an understandable point, although one could argue that customers also pay for all of these, and they often populate the translation memory with the contributions of many translators and not just one, as well as perhaps running numerous quality verification checks; therefore, it seems logical that they also want to benefit from the use of tools. There were five translators that thought it was "Unfair" (Translators 8, 9,

11, 12 and 19). Four of these five were not satisfied in the previous question about pricing. Translator 8 thought the pricing was “Unfair” in this particular case, although she was satisfied with the pricing in general. She mentioned, “most of the time, fuzzy matches need as much work as no matches”. Fifer also reports on the differences in the ranking of Fuzzy matches and human judgment (2007). Translator 5 thought that the payment was “advantageous”. Incidentally, she was also “Highly satisfied” with her work.

How do you revise fuzzy matches when working in SDL Trados or similar tool? (You need to select one option per row.) After downloading a segment...						
Answer Options	Never	Rarely	Sometimes	Frequently	Always	Response Count
<i>I read the Source, then correct the Target segment.</i>	0	3	2	10	9	24
<i>I read the Target, then the Source segment, then I make the changes.</i>	4	6	8	5	1	24
<i>I look at the changes marked by the tool, then I correct the Target segment.</i>	2	6	3	10	3	24
<i>I read the Target, then I look at changes marked by the tool, then I correct the Target.</i>	5	8	5	4	2	24
<i>Other (please specify)</i>						2

Table 76: Question 8: Fuzzy match revision

Regarding the methodology for revising Fuzzy match segments it seems that the most common practice in the translators’ opinions is to download the segment, read the source and then correct the target while also looking at the changes marked by the translation tool. Obviously, the table shows that the translators report a combination of methods, but it appears to be less common to just focus on the target texts. Another interesting point is that there is quite a spread in the option “I look at the changes marked by the tool, then I correct the Target segment” as we would have imagined that almost all translators would select “Always” in this option. However, it seems from these responses that reading the Source first and then the Target is more frequent.

How adequate is the payment of proposed matches in Machine Translation in relation to the productivity you obtain with them?							
Answer Options	Highly unfair	Unfair	Fair	Advantageous	Highly lucrative	Rating Average	Response Count
<i>I think it is</i>	1	8	10	0	0	2.47	19
<i>Please, explain how you have been paid to post-edit so far.</i>							19

Table 77: Question 9: MT pricing

For this question, the translators were asked to explain how they had been paid to post-edit so far. The rating average for this question is 2.47, lower than the other two questions that refer to payment (general pricing and Fuzzy match payment). There are

19 responses to these questions because, as we saw above, six translators declared not to have had experience in post-editing MT output. However, one of these six translators (Translator 22) responded “Unfair” to this question because “I don’t like the idea of Machine Translation”. Of the 18 translators with experience, ten think the payment is “Fair”. Translator 13 thinks that the payment is “Highly unfair” (as the fuzzy match payment) but she was “Satisfied” with payment in general. Perhaps, she carries out other types of work where she is satisfied and post-editing is only a small part of her work (1-25 percent according to her response). Seven translators think that it is “Unfair” (Translators 7, 12, 15, 18, 19, 20, 22 and 24). They are normally paid “per word but at a higher rate than editing human translations”, “the same rate as when translating with CAT tools”, it is “dependent on the customer”, “70 percent of the word rate” and “50 percent of the word rate” or “the same rate as proofreading”. Translator 15 commented:

It depends on the client/project. Many times quality expected from postediting is the same as from human translation (this goes against the idea of postediting, btw). Some clients will ask me to take lots of things into account (terminology, style, etc.) when postediting so in the long run it is not cost effective to me.

The ten translators that think the payment is “Fair” commented that they were paid something “between the no match rate and revision rate”; “70 percent of the word rate” (Translator 8 said that it was fair because she doubles her productivity in this type of projects); “full rate”; “per hour with an agreed productivity rate reflecting the real time the task takes”; “based on the quality of the MT output”, or they were paid a rate “corresponding to a high fuzzy match” (it was then fair when the MT output was “good” but not so when the MT output was “poor”).

What would be, in your opinion, the ideal payment method for post-editing MT output?		
Answer Options	Response Percent	Response Count
<i>Per word</i>	41.7%	10
<i>Per hour (with an agreed productivity rate)</i>	50.0%	12
<i>Other</i>	8.3%	2
<i>If you chose other, please specify...</i>		2

Table 78: Question 10: Ideal payment method for MT output

There is no clear answer on the ideal payment method. The two translators that selected “Other” summarized this dichotomy. Translator 15 says that per hour should be more appropriate if all requirements and expected quality are defined at the start of the project. Translator 22 mentions that it depends on the MT engine and output, and she

finds there is no “universal solution”: both per word and per hour can work together but each project should be treated individually.

Do you like using MT as part of the localization process?							
Answer Options	Strongly dislike it	Dislike it	Am Indifferent	Like it a little	Like it very much	Rating Average	Response Count
<i>I...</i>	5	4	8	5	2	2.79	24
<i>Please, tell us why</i>							20

Table 79: Question 11: Predisposition to MT

The average is 2.79 out of 5. Five translators “Strongly dislike it” (Translators 16, 20, 21, 22 and 24), coincidentally three out of these five translators (Translators 16, 21 and 22) declared not to post-edit, so it seems natural that translators who dislike the task will not accept this type of work. During the project, however, these translators did show productivity increases when working with MT, but of course this does not mean that they were actually “enjoying” it. Four said that they “Dislike it” (Translators 4, 5, 17 and 18). Again, Translator 17 had declared that she did not post-edit, so this seems quite natural. Translator 5 dislikes MT because she says her productivity goes down when using it. In this particular project, Translator 5 did increase her productivity, but of course this is not applicable to all of her post-editing projects and we know that in this particular case the quality of the output was high. Translator 4 commented, “Some segments are disastrous. I think this will improve with time”. Eight were indifferent (Translators 2, 3, 6, 7, 13, 14 and 23) although two of these had no experience in post-editing (Translators 3 and 6). Translator 14 made an interesting comment:

I think MT is marked by current business trends. As a translator, I evaluate each task proposed in term of time and rate, and if I agree on the job proposed, I accept it. MT is a new tool and as a professional I should be acquainted with it so as not to be out of date.

Translator 23 mentioned, “I have to review the translation given by the machine the same way I review my own translation, so it is fine with me”. There were seven translators (Translators 1, 8, 9, 10, 11, 12 and 19) that either “Like it a little” or “Like it very much”, and they gave the following reasons: they can leverage content; it is dynamic and “physically advantageous” as it avoids having to type continuously; it helps consistency and the translation of repetitions (especially if the translation memories used to train the engine are well maintained); it increases productivity; it helps accuracy; and it is useful especially in texts with similar patterns.

7.2. Reviewers' opinions

There were only three reviewers on this project but we thought it was still important to know their opinions about the work they do, since they are not only reviewers but also professional translators. These were the characteristics of this group of reviewers:

- More than eight years' experience in localization;
- More than eight years' experience using tools;
- Between two and eight years' experience in localizing business intelligence software;
- Experience working for MicroStrategy;
- Between two and six years' experience in post-editing;
- More than four years' experience in reviewing;
- Daily throughput of 5,100 to 7,000 words (according to their responses).

Judging from these characteristics, they are a group of very experienced translators. Because there are very few responses in the questionnaire, we will not present each table of results; we will simply describe their responses according to the different themes.

7.2.1. Review methodology

When they were asked how they review their own work, the reviewers gave different responses. They review after finishing one segment, a file or a batch of files. They hardly ever review their work after they have completed a days' work. On the other hand, to review others, the most common method was to read the source, then check the proposal from the tool, and then implement corrections in the target (for Reviewers 1 and 3) and read the target, then the source and then implement corrections (for Reviewer 2). If they were dealing with fuzzy matches, all three reviewers read the source text and then implement the changes in the target text; two reviewers check the changes marked by the tool (Reviewer 2 does not tend to do this). All of them use either the LISA QA form or other proprietary forms (from their customers). They do not use J2450 and they always use some kind of form to report on the quality of the translations.

7.2.2. Pricing

Reviewer 2 is satisfied with the price paid for reviewing but Reviewers 1 and 3 are not, because the price is low considering that they also have to fill in the QA forms and that they have a greater responsibility for the translation quality. Coincidentally, Reviewer 2 was “Very satisfied” with the work she does as a reviewer, while Reviewers 1 and 3 were just “Satisfied” if the price was not considered. Perhaps, the price received is influential for Reviewers 1 and 3.

Reviewers 1 and 3 think the price paid for fuzzy matches is adequate although Reviewer 1 mentioned that they are paid according to a fuzzy match rate despite the fact that the TM is not the only reference they have to consult, and this might make the payment unprofitable if there is a high number of reference material they need to check for each segment. Reviewer 2 finds the price paid “Unfair” due to file formats and tagging in files.

Reviewer 3 finds the payment of proposed MT match segments Fair but she mentions that this depends on the quality of the output, so she states that it is not always “Fair”. Reviewers 1 and 2 found it “Unfair”. Reviewer 1 mentioned that this is due to incorrect terminology and excessive tagging. Reviewers 3 and 2 think that it is better to get paid per word for the post-editing task, while Reviewer 1 thinks it is better per hour with a productivity rate agreed upon.

7.2.3. Opinions of MT

Reviewers 1 and 3 dislike working with MT output. Reviewer 1 dislikes it:

[...] unless it is a very simply structured document without specific vocabulary and without tags, it normally takes more time to post-edit it than from translating from scratch, and you are discounted a significant percent of the fee.

Reviewer 3 dislikes it mainly because:

[...] we cannot predict how useful it is going to be in the end. Sometimes it makes things much [more] difficult than translating without it. I cannot trust in [sic] the MT output, and the most time-demanding task in the process, that is, consulting the reference material, glossaries, etc. is still part of the job. In few occasions [sic], we have been told the MT is trustworthy and it actually was.

Reviewer 2 is indifferent.

7.3. Translators' opinions of the assignment

The final part of the questionnaire contained questions about this particular assignment. The objective was to see what translators thought of the tool, the methodology and the segments proposed, and to observe if these answers could explain some of the results. As we did in the previous section, we will present the questions with the results and a brief analysis.

How easy to use was the tool employed in this assignment?							
Answer Options	Very difficult to use	Difficult to use	Average	Easy to use	Very easy to use	Rating Average	Response Count
<i>It was</i>	1	1	2	6	14	4.29	24
<i>Comments</i>							5

Table 80: Question 1: Opinions of the on-line post-editing tool

The rating average is 4.29 out of 5, which is quite positive. Two translators found the tool either Very difficult (Translator 13) or Difficult to use (Translator 24) because it was too slow (this presumably refers to the bandwidth that Translator 13 had, she could not perform the interview using Skype for this same reason), it did not allow the translator to go back and check the segments, the source could not be copied, it did not allow one to search similar strings or access terminology.

Was the tool used in this assignment similar to the tools you normally use?							
Answer Options	Very different	Different	Same	Similar	Very similar	Rating Average	Response Count
<i>It was</i>	2	12	0	9	1	2.79	24
<i>Comments</i>							6

Table 81: Question 2: Similarity with other tools

The rating average is 2.79, so opinions were divided on this question. Translator 1 and Translator 13 said it was “Very different”; Translator 11 thought it was “Very similar”. Possibly the explanation for these different opinions is that there were basic functions that were similar but this tool lacks common functions found in other tools. So, depending on the point of view, the tool could be, in fact, different or similar.

How useful were the proposed segments you edited?							
Answer Options	Not useful at all	Not really useful	Indifferent	Useful	Very useful	Rating Average	Response Count
<i>They were</i>	0	0	7	17	0	3.71	24
<i>Comments</i>							5

Table 82: Question 3: Usefulness of MT matches

The rating is 3.71. Note that 17 translators found the proposed segments “Useful”, which corresponds to the results seen in the Quantitative part, where all translators showed a productivity increase when using MT or Fuzzy matches and also smaller number of errors. Five translators commented that some segments were useful and others were not, which is quite understandable as we saw with the TER score that some segments “needed” few edits while others required a substantial number of edits. This is also reflected in the different times and numbers of errors per segment.

How comfortable was the proposed review method (reviewing immediately after translating/editing a segment)?							
Answer Options	Very uncomfortable	Uncomfortable	Indifferent	Comfortable	Very comfortable	Rating Average	Response Count
<i>It was</i>	1	6	3	8	6	3.50	24
<i>Comments</i>							7

Table 83: Question 4: Review method

The rating average is 3.5. This is interesting, considering previous responses about the tool, but it is in line with the opinions about the review methodology, where most translators stated that they reviewed each segment after completion. Note that 17 translators seem to be either indifferent or comfortable with the review methodology. One translator is “Very uncomfortable” and this is Translator 13. As we saw above, she has a different revision methodology and is also uncomfortable with the tool, and she made a high number of errors (although she had the highest speed). So being uncomfortable with the review method could be another reason for her poor performance in terms of quality. It could also be that Translator 13 simply went over the exercise too fast and did not pay enough attention to details. However, it is obvious that this participant was uncomfortable with the task. There were six translators that said they were uncomfortable (Translators 2, 6, 16, 17, 21 and 24) because they could not go back to the segments after completion or because they were not used to checking the segments after completion. Translator 24 offers an interesting comment:

Immediate reviewing demands extra attention in two tasks (translation and reviewing) for each segment and that can be exhausting. I prefer working in one task at a time (first translation; after finishing, reviewing) simply because at the end of a translation you have a greater knowledge on the subject and style, and you are able to do a better review.

In the case of the other six translators, there is no clear relation between the data we gathered in the project (in terms of speed and errors) and their responses. This is not because they could have performed better in another environment (it is impossible for us to know that) but because they did not appear to perform better or worse than other translators who declared they were “comfortable” with the tool.

How productive were you in this assignment in comparison to regular human translation (without any aid)?		
Answer Options	Response Percent	Response Count
Equal	25.0%	6
Faster	45.8%	11
Slower	20.8%	5
I don't know	8.3%	2

Table 84: Question 5: Perceived productivity

Most translators found that they were either faster or equal (70.8 percent) in terms of productivity. This corresponds partially to the results obtained during the assignment, if we compare the speed obtained when processing MT and Fuzzy matches with the human translation (No match segments) in the same assignment. Note that 20.8 percent said they were slower (Translators 1, 13, 20, 22 and 24). It could be that the overall exercise was slower for these translators in comparison to their own personal work, despite showing an increase in productivity with regards to the No match segments if mean values are considered in this particular project. It could also be that they merely *perceive* they were slower, when in reality they showed an increase in productivity with respect to the No match segments.

During the assignment, how did you check terminology in the proposed segments? (You will need to select an option in each row.)						
Answer Options	Never	Rarely	Sometimes	Frequently	Always	Response Count
<i>I accepted the terminology proposed.</i>	7	3	5	3	6	24
<i>I checked each term in the glossary provided.</i>	0	0	0	4	20	24
<i>I checked each term in the glossary provided and I did some research on my own.</i>	0	5	14	2	3	24
<i>I checked those terms that did not seem appropriate in the proposed segments in the glossary provided.</i>	2	2	4	2	14	24
<i>Other (please specify)</i>						3

Table 85: Question 6: Terminology

In our previous project (Guerberof 2008) we observed that terminology was a problem for the participants in terms of number of errors after post-editing and this is the reason why this question was included. As we saw in Chapter 5, terminology was also an issue in this project. Twenty translators responded that they “Always” checked the terminology in the glossary and four that they “frequently” checked term in the glossary. However, according to the three reviewers terminology was a problem in the target translations (Terminological errors were the second highest type of errors in the project). The translators that responded “Never” to this option (Translators 2, 4, 7, 11, 17, 20 and 23) were among the ones with fewer terminological errors. Translator 10, with the second highest number of Terminology errors answered “Always” to checking each term in the glossary, doing research on certain terms, and checking the terms that did not seem appropriate. Translator 19, with the highest number of Terminology errors, responded “Always” to “I accepted the terminology proposed”, and this might explain the high number of errors since the instructions indicated that the Glossary had priority over the proposed texts. Translator 15 with one of the lowest number of Terminology errors and the lowest number of errors overall, commented that he exported the glossary to a tabbed txt file and used Xbench (a free tool that offers quality assurance features developed by ApSIC, a language service provider ²) for quick searches and pasting the results back in the interface. Perhaps, terminology tools can help translators to offer better quality. Translator 20, on the other hand, commented that she had methodically checked the glossary and she had good results in terminology and in the global number of errors. Both of these translators were in Speed group 3 (the fastest), whereas they had a low number of errors, and of Terminology errors in particular.

At the end of the questionnaire, the translators were asked to add any comment that they might deem relevant. They commented that post-editing was not just revision, that the type of text had a direct impact on the efficiency of post-editing, that some segments contain the wrong word order for Spanish (presumably the MT proposals), that they found the tool interesting for technical texts, that the task was easy and the reference material useful. There were comments on improvements to the tool, for example, to include a Concordance feature, to have the source text copied automatically for the No match segments or to be able to review segments after clicking Next.

² http://www.apsic.com/en/products_xbench.html

7.4. Reviewers' opinions of the assignment

We wanted to know how the reviewers perceived the quality of the translations immediately after completing the task, when the texts would be “fresh” in their minds. We included some questions on their opinions about this particular assignment (see Appendix D). We summarize their responses below.

7.4.1. Translation quality

All three reviewers agreed that the quality of the translations reviewed (considering the number of errors) was “Average”. This is understandable since the range went from 44 errors to 167 errors (aggregated value of the three reviewers). Reviewer 2 commented that she had noticed in the text:

...lack of fluency, unnatural language and misuse of articles and prepositions due to ambiguity of some sentences that led translators to implement ambiguity too in the translation.

As happens in “real” projects, sometimes source sentences are difficult to decipher and translators interpret the text differently. Reviewer 3 thought that Translator 12 did a particularly good job. This was the translator with the fourth lowest number of errors according to the three reviewers, and second for Reviewer 3 (Translator 15 ranked first). Translator 12 was in Speed group 1 (the slowest). We saw that speed was not a factor when considering the number of errors, and this is clear for those translators that had fewer errors (some were in Speed group 3, the fastest, and others in group 1, the slowest). Reviewers 1 and 2 just responded that three or four translators had done a good job, following the instructions and the glossary provided, and they provided excellent language quality, but they failed to name those translators. Reviewer 1 marked fewer errors for Translators 20, 6, 8 and 11, while Reviewer 2 marked fewer for errors for Translators 8, 15, and 20. They might have been referring to these translators. Reviewer 3 mentioned that Translator 10 was particularly poor because the glossary was not consulted. Translator 10 is the translator with the most aggregated errors and also with the most errors according to Reviewer 3. Also, Translator 10 did not follow the instructions correctly at the beginning of the assignment (see section 3.5). Reviewers 1 and 2 also responded that there were two or three translators that were particularly poor because they had a high number of spelling mistakes, they produced unnatural/awkward

sentences, they did not use the glossary and they did not read sentences before confirming them. The translators with most mistakes for Reviewer 1 are Translators 19, 18 and 10; and for Reviewer 2, Translators 19, 3, 18 and 13.

7.4.2. Difficulties in the text

The reviewers were asked if there were segments that they had found particularly difficult to translate. Reviewer 3 mentioned that in segment 78 “the display style and color of the gauge faceplate, gauge border, and needle” was difficult to translate as it was unclear if the source meant “the color and the display style” or “the display color and the display style”. This segment had the second highest number of errors for the three reviewers. The segment also has an average TER of 57.29, which means that translators made a considerable number of edits, the mean speed value was 7.11 words per minute, one of the lowest mean speed values of all segments. Reviewers 1 and 2 do not give details of specific segments but they mentioned that there were segments that were difficult to translate. Reviewer 1 mentions that “the wording was ambiguous/not very clear and many translators mistranslated them”. Reviewer 2 explains that the “subordinate clauses led to the above mentioned ambiguity. It is not clear which noun relates to the clause”.

When asked if they thought the proposed segments were useful for the translators, they answered that they were either “Useful” or “Very useful”, but Reviewer 1 commented that:

Fuzzy matches were in general useful, still many had problems (grammar/vocabulary mistakes) and many translators did it better with no matches than with fuzzy matches and with MT. MT segments were normally unnatural and not very useful.

This was really surprising, since Reviewer 1 found 309 errors in No match, 187 errors in Fuzzy match and 171 errors in MT match segments, and the reviewers could see the origin of the segments while reviewing. This might indicate that one of the issues with machine translated segments, and indeed with all translations, is that if one segment is very poor there is a tendency to generalize to all segments. However, if we look at particular translators, some had more errors in MT than in Fuzzy matches if the difference in words is not considered. As we saw in the TER section, MT segments tended to be more “extreme” than Fuzzy match segments, meaning that they require both many or few edits, and this might have caused Reviewer 1 to believe that these MT

segments were not as useful. Nevertheless, all three reviewers found the segments useful in general, and this is reflected in the percentage of segments with errors that they marked.

7.4.3. The LISA form

Reviewers 1 and 3 found the form “Easy to use”, and Reviewer 2 found it “Very easy to use”, although she commented that it was difficult to distinguish between Accuracy and Mistranslation errors. It seems that this was indeed the case, as she classified most of the errors (188) as Accuracy errors, and only six as Mistranslation errors, while Reviewer 1 had 86 Accuracy errors and 55 Mistranslation errors, and Reviewer 3 had 37 Accuracy errors and 140 Mistranslation errors. The reviewers did not have the same number of errors in the other categories, and the difference in the number of errors in those other categories was not as pronounced. Surprisingly, however, the three reviewers responded that they would not change the categories in the LISA form and that they found the severity scale either “Easy” or “Very easy” to use. Reviewer 1 mentioned that “Still, sometimes it can be a bit subjective determining [sic] the border from minor to major and from major to critical, but overall, it's easy to use.” We believe, however, that the current classification and the severity levels are not that clear for reviewers, or simply that errors are difficult to classify in general. After the project was finished we had to contact the reviewers to clarify certain error classifications or error counts, and they had to redeliver the file twice with amendments (see section 3.5).

7.4.4. Assignment review method

The three reviewers had different opinions of how comfortable the review method was. Reviewer 1 thought it was Comfortable. Reviewer 2 thought it was Very comfortable and she explained this was because in a normal review she did not use track changes, but she had to copy every error (source text, target text and her proposal) in the given review form. Reviewer 3 thought it was Uncomfortable because she is used to reviewing with a translation memory.

When asked how they checked terminology, the three reviewers responded that they always checked each term in the glossary provided. Reviewers 1 and 3 checked the terms in the glossary and did some research on their own. Reviewer 1 also checked those terms that did not seem appropriate.

Chapter 8: Debriefings

In this chapter, we present the results of the debriefings carried out with the translators and reviewers. As we mentioned in section 3.4.4.3, we carried out informal semi-structured interviews immediately after the assignment was completed in order to capture translators' and reviewers' opinions, feelings, perceptions about the assignment, machine translation, translation process, and any other data that might emerge. We were looking to validate or expand on the data gathered through the questionnaire. The debriefings tended to last 15 minutes or less with the translators, and 30 minutes or less with the reviewers.

8.1. Translators' debriefings

We interviewed 19 translators (Translators 2, 3, 10, 13 and 16 were not available for the interview for different reasons as we explained in section 3.5) and the three reviewers involved in the project. We translated and transcribed all recordings into English as can be seen in Appendix I and in Appendix J. For the translation and transcription, our objective was to capture what participants said during the debriefing in the same way they had expressed themselves in Spanish, and therefore we fully translated and transcribed them. At the same time, an effort was made for the text to be understood so certain repetitions were not rendered or certain structures were clarified, but this was kept to a minimum. The interviewees were expressing themselves in their native language and from their own homes or offices, so they were very relaxed and open to the questions - an exchange of ideas could take place. None of the participants declared having a problem with being recorded. The recording applications for Skype that we have used could record 15 minutes at a time, and therefore if the interview went on for longer, we had to stop and start a new file (this is indicated appropriately in the transcripts). The Spanish recordings are also available in mp3 format.

The same questions in the same order were put to all interviewees, although on occasions the wording was slightly different to add to the fluency of the conversation. The questions were:

- What did you think of the instructions for the task, including the glossary?

- Did you know in advance that this was a project containing MT segments? Did you think about it at any particular moment during the assignment? How did you know?
- Did you notice any difference between the proposed segments? Do you have any examples?
- Was there any segment that you found more difficult to translate or edit? Why?
- Which questions in the questionnaire were most difficult to answer?
- How did you feel doing the task?
- Would you like to add any comment?

Once we had all the debriefings translated and transcribed, we used NVivo 9.0 to collect and analyze the data. We had an initial framework as described in section 3.4.4.3 created from the type of questions we were asking. Once we had all the debriefings, we manually coded the questions and they were organized according to this initial structure and we modified it slightly to this final framework:

Integration of TM and MT

1. Assignment

1.1. Instructions

1.1.1. Unclear passages

1.1.2. Clarity

1.2. Glossary

1.2.1. Issues

1.2.2. Clarity

1.2.3. Completeness

1.2.4. Solutions

1.3. Questionnaire

1.3.1. Difficult

1.3.2. Easy

1.4. Tool

1.4.1. Advantages

1.4.2. Disadvantages

1.5. Segments

1.5.1. Awareness of MT

1.5.2. Type of segments

1.5.3. Difficulties

- 1.5.4. Quality
- 2. Feelings
 - 2.1.1. Dislike
 - 2.1.2. Like
 - 2.1.3. Neutral
- 3. Machine Translation
 - 3.1. Opinions about MT
 - 3.1.1. Positive
 - 3.1.2. Negative
 - 3.1.3. Mixed
 - 3.2. Knowledge of MT processes

Since the questions dealt primarily with the assignment and the objective was to set in motion a conversation in Spanish, we are not going to present a quantitative analysis of references to these topics but rather a *qualitative* account of what the translators said during the debriefings.

8.1.1. Assignment

This section includes all coded data for the assignment rather than general opinions or feelings about MT or the profession.

8.1.1.1. Instructions

All the translators interviewed found that the instructions were clear. There are 25 references from 19 sources (the 19 translators that were interviewed) mentioning that the instructions were “clear”, “very clear”, “I had no problems”, “simple” or “concise”; some even mentioned that they were pleased that the instructions were “short and clear that is what you normally want”, and “all that was very clear and good because it was short, brief”. Translator 24 gave an insightful view of what translators normally face in terms of instructions and why they prefer instructions that are short and clear:

Well, in many cases you are sent instructions that can occupy pages and pages and you have to read... and that it isn't very economical especially if the job is short. In general I think that as translators we prefer instructions that are clear and concise and then if there are problems during the process to have good communication with the manager to solve them.

There were, however, unclear passages that caused confusion among the translators. Although there were only four translators that made comments on these unclear passages (Translators 11, 18, 20 and 23), we believe that other translators might have made decisions based on these unclear instructions (perhaps based on their own previous experience) that might have resulted in having more final errors (according to the reviewers). One aspect that was unclear was that translators did not know that there was machine-translated text in the assignment (we have explained the reasons for this in section 3.4), so this meant that when they saw the instructions “You do not need to introduce preferential changes, just correct errors, and you do not need to re-write the text in a certain way if it is not to correct an error, you do not need to insert changes to improve the text” (see Appendix C), some were unsure as to what this meant. Translator 11 said “when I started working on the segments I realized what the instructions meant by strictly correct identifiable errors”. This translator sent changes to his translation by email when he realized his errors (see section 3.5). Some translators were not prepared to face word order problems, for example, and this could have resulted in accepting a proposal that sounded correct but that it was in fact wrong. Another aspect was the dichotomy between the quality expected (publishable) and the instruction not to make preferential changes: “You see segments that are translated correctly but they don’t adapt exactly to the style”, as Translator 20 said. The instructions told translators: “Do not introduce preferential changes, only correct errors or make changes that you are certain about and that are fully justifiable” and “full accuracy and no mistranslations with regards to the English text, compliance to Spanish language rules of grammar and spelling, compliance to the terminology following the glossary provided (Glossary.xls), and compliance to style according to the instructions explained below” (see Appendix C). We perceive certain confusion among translators as to the differences between “preferential” and “style”. As defined in the instructions, “preferential” referred to changes that do not correct errors that are justifiable, and “style” referred to those segments that did not follow the Style guidelines specified in the instructions. We agree, however, that the best way to show the type of errors to be corrected is to include specific examples; we also know from experience that on occasions the examples themselves can cause confusion. In brief, instructions on quality expected and type of desirable changes need to be extremely precise and to include examples.

Another aspect that caused confusion for Translator 18 (as we have seen this translator had the third highest number of errors and he reported that he did not

translate) was the instruction “All software options will be translated in Upper Case as in the Source English text”. He was unsure if the whole nouns in the software option were to be in upper case or only the initial letter of that software option. He only understood this at the end of the assignment but he could not go back and fix the other software options. This instruction was ambiguous. However, since the glossary was to be the primary reference and in this glossary only the initial letters of software options were in upper case in Spanish, we believe that Translator 18 perhaps did not consult the glossary as much as was desirable, despite stating in the questionnaire that he “Always” checked each term in the glossary.

Finally the instruction “The infinitive in English will be translated as infinitive in Spanish” caused problems, mainly because of the lack of context when translating the segments. On occasions the translators did not know if they were dealing with an infinitive in English or simply a prepositional phrase “To add...”, as in a procedural document. These are issues that arise from translating individual segments without context. The translator does not know if that particular segment is going to be a title, an instruction or a description.

8.1.1.2. Glossary

Most translators (18 references from 18 sources) refer to the glossary as “good”, “complete” or “very complete”, “simple”, “practical”, “fairly consistent”, “correct” or “fairly populated”. The glossary contained most of the vocabulary and all the software options. This was a positive aspect of the assignment because, as Translator 9 pointed out, “many times you look for many things that are not in the glossaries, or in the translation memories that are sent to you by some clients”.

As with the instructions, there were nevertheless, some aspects that made the glossary uncomfortable for some translators. We found 14 references to this from 11 translators. One aspect was the fact that the glossary was complete but perhaps too detailed. For example, Translator 12, with low number of errors, mentioned that:

Although at the beginning I saw some [words] and I didn’t look at the glossary that much because they are words that you have always known, but well, for example, I realized that for “add” I had used *añadir* in a given moment but then in the glossary I saw *agregar*.

The fact that there were more words than expected and that translators could not go back and review the file meant that on occasions a translation was used that, although perhaps a standard translation, was not in the glossary. The translators stated in the questionnaire that they checked each term in the glossary, the fact, however, is that some terms might appear too obvious to check. Of course, it is impossible to know in advance what terms translators will or will not know, and in real-world practice, experience with a particular customer tends to mean that not all terms have to be checked in the glossary.

Another aspect was that certain translators did not agree with how the terms were translated. They stated that some terms were missing from the glossary or that other customers (such as IBM or Microsoft, companies that create industry trends in terminology) translated certain terms differently. However, these were more like general comments than actual shortcomings of the glossary.

The aspect that caused more headaches, judging by the number of comments (six) was that the glossary was not integrated into the tool and it had to be consulted externally in an Excel file. Four translators, however, found solutions around this. Translator 12 simply checked every term in the glossary and Translators 11, 15 and 23 used Xbench to simplify the searches.

8.1.1.3. Questionnaire

We asked the translators which aspects of the questionnaire they found difficult to answer, if any. Fifteen translators gave their feedback on the aspects they found most problematic to answer. However, eight translators mentioned that “it was very easy to complete”, “I didn’t have major difficulties”, “the questions were clear”, and “they were not difficult to respond to”. Four out of these eight did not report any difficulties. The main issue translators had when completing the questionnaire was answering the questions related to rates, since “sometimes you are more satisfied than other times”, “it is relative because not all the jobs in machine translation take the same amount of time”. In brief, translators reported that the question was too general, given that satisfaction with rates depends on customers, subject matter, nature of task (if it involves machine translation, for example), actual rates and the financial situation in a given country. Translator 17 explained some of the complexities regarding rates paid:

[...] it is very difficult to generalize on this topic [...] I work for several translation agencies and sometimes for publishing houses with different rates [...] in Argentina

there is a range of prices that is huge and in general it depends on how many intermediaries there are along the way. [...] sometimes there are a lot of intermediaries, and what arrives to the freelancer is a low rate...

Another difficult question was the one referring to how they review their translations (see Table 76 on page 204). Some translators found this difficult to define as it depended on the peculiarities of the projects, and others were not aware of the revision method used. Translator 18 explained “Sometimes all [the translation process] is so automatic that I don’t reflect on the actual process”. We saw, however, that this translator did have problems in the whole process (high number of errors, not understanding instructions and not following the glossary).

Some translators that had not performed post-editing tasks previously found those questions difficult to answer, understandably so (in fact, they were asked to answer them only if they had had experience). Translator 5 did not know her typing speed so she found this question difficult to answer. She did take, however, a small test in Word, controlling her speed with a timer. Translator 24 mentioned that he found it difficult to say if more effort was required to revise human translations or post-edited material (he responded that more effort was required to post-edit).

8.1.1.4. The tool used in the assignment

Although we did not ask specific questions about the tool, translators referred to certain advantages and disadvantages that the system used for the project had. There were seven translators that mentioned that the tool was “very practical” and “easy to use”. Translator 7 found that the fact that translators could not go back and make changes in the tool made the task more agile, “It was as if you had to do each segment in the most efficient way possible.” There were, however, five translators that commented on certain disadvantages of the tool: not being able to identify the words requiring changes in each segment (as other CAT tools do), not having a spell checker, not having the English source copied by default in the target segment in the No match category, and not being able to go back and correct a processed segment.

It seems that translators like the fact that the tool was simple and easy to use. There were few options available in the tool and the screen was not cluttered with different windows. However, they were looking for options that they have grown accustomed to and that facilitate their work with the tools that they are currently using such as Concordance feature or an integrated glossary.

8.1.1.5. Segments

There were three questions in the debriefings directly related to the segments: one was related to their awareness of having MT segments, another to the differences in the segments in terms of type of edits, and the third to difficulties found in certain segments during the task. Thirteen translators realized that MT output was involved in the task, either when they received the instructions or when they started working on the assignment. Before starting the task, five translators knew or strongly suspected that there would be MT output. They imagined this when they received the emails to participate in the project, despite the fact that it was not mentioned, because the email said there would be Fuzzy matches (see Appendix B). Although Translators 12 and 15 were among this group of translators and they performed very well in terms of errors, the other three translators had an average or poor performance, so we cannot suggest that knowing a priori the exact nature of the task was an advantage or that it resulted in better quality results. Still, with the instructions and during the task, they all became fully aware, mainly because of the “changed structures”, “expressions where it was clear they had not been translated by a person”, the “very literal translation”, or “word order” and the fact that the instructions mentioned not to correct only errors that were certain about and “not style issues”.

Translator 17 was not aware that the project involved MT, as she explains:

Really, the perception that I had was not that it was machine translation but that it was a translation memory. Except some segments that clearly, well I imagined they were modified to see the correction made or if the error was noticed.

This is due to the overall high quality of the MT output, as we have seen in the Quantitative part. Notwithstanding, some segments were poor. Translator 17 was not fully aware that she was dealing with MT and thought the segments were seeded with errors. Still she had a low number of errors (58 in total, aggregated value from three Reviewers) and she was in Speed group 3 (the fastest group). Fifteen translators perceived differences between the segments; six clearly stated that there were Fuzzy and MT matches. Translator 11 even thought that he was able to identify the type of match:

[...] I noticed the difference was mainly similar to any fuzzy segment in any translation memory that shows, for example, a segment with a 95% match. Everything is the same

but there is one word or one section of the document that does not coincide with it, precisely because the update was not done with respect to the new source.

The others perceived differences in terms of quality of the segments: some were very good, and others were poor, although they specified that the poor quality was not in the majority of the cases. Some translators seemed to imply that those segments that were “very good” “belong to a human”, as Translator 22 explained:

There were some that were very good, you could say that they belong to a human, but others were obviously from a machine.

Others mentioned that if this was MT output then it was “very good”. It is interesting to see that some translators assumed that the proposed texts that contained fewer errors were human translations when in fact many MT segments were not changed by translators (see section 4.5) because they were of acceptable quality. A similar experience reported in He et al. (2010b) where post-editors mistake MT outputs for TM outputs. We have no way of knowing how translators perceived each individual segment and whether the segments they thought were human were indeed human and not MT. The majority of translators in this group were familiar with certain type of MT errors (word order, wrong structure) and they appeared to be able to identify these segments very quickly but when it came to segments that flowed well they might have assumed that these were human translations. The other four translators did not think there was a clear difference between the segments, although they might have thought that some were better than others in terms of linguistic quality.

Seven translators made reference to the overall quality of the segments: some mentioned that it was “pretty good”, “fairly acceptable”, “very good quality”, “fairly good”, “a high percentage of what was already translated was better than what I expected, really”, another mentioned that, “segments that were longer and more complex; you had to almost completely change them, they were not the majority”, or that “I had to read again the source and then rewrite, reformulate practically the whole sentence... and there were also many segments that were perfect.”

8.1.2. Feelings

The translators were asked how they felt during the task. Ten stated that they had liked the task because “it was interesting”, “because it isn’t what I normally do”, “it was

something enjoyable”, “I liked it quite a lot”, “I was positively surprised” (especially with regard to their perceived quality of the MT), “I like the tool” and that it was “dynamic”. Translator 4 was particularly pleased for the following reasons:

To be honest, very good. It was a pleasure. I’m very used to this type of translations, all that is software, etcetera. And to be honest the task was very good, mainly because we had all the material. If there was anything to consult, any terminology to consult and it was not in the glossary, I took the Microsoft terminology databases as terminological reference but we had all the tools to be able to do it, and the instructions were very clear.

The translators seemed to be pleased that the task was short, uncomplicated and at the same time it was outside their normal routine. They also found it interesting because they felt they were involved in a research project that involved acquiring knowledge for the profession. Seven translators were quite neutral in their comments. Some felt it was another job, “like a normal project, like the ones that are normally done for example with Trados”, “it is very similar to what I do as a professional”, “I felt comfortable because it was very similar”. Others felt they had not experienced any particular problem; and still others thought that although they had to pay more attention: “it was not that horrible”.

Finally, Translators 21 and 22 did not like that the task because they found it either tedious or they did not like working with machine translation. In their own words:

A translation of this type, I found it tedious, it takes a long time and in the end you end up retranslating almost everything in the end, at least in my case. I haven’t done it a lot, but the times I have done it, this was my experience, no, no, I wouldn’t like to do this daily and with large projects. Yes, I think it was tedious. (Translator 21)

A bit uncomfortable, because I insist, I don’t like working with *machine translation* [English in original]. Even if this one I could tell it was of very good quality, but no, it isn’t something that I like. (Translator 22)

Translator 22 also mentioned in the questionnaire, “One never can trust 100% [sic] on a Machine Translation”. It is interesting that these two translators also had quite a high number of errors (Translator 21 had 80 errors in total and Translator 22 had 98) and that they had fewer errors in Fuzzy matches than in the other two categories: MT and No match. They belonged to Cluster 3, that is, translators with no post-editing

experience. It would be interesting to know if not liking the task had any influence on their performance, but of course, we would need to test them doing a task they liked doing, for comparison purposes, and that was beyond the scope of this project. What we can say is that other translators that *did* like the task had a similar number of errors or equal speed. Therefore, in this particular project, we cannot establish any correlation in this respect.

Eight translators made reference to the fact that the task was easy. Translator 14 mentioned that “it was the ideal translation task” because the job was clear and the reference material was at hand, and that she “had never worked with a tool that was this easy”. A lot of these comments made reference to the fact that in a localization project nowadays translators need to install new software continuously, learn how it works, work with different tools at the same time, and have an array of reference materials opened. All of these additional activities seem to cause a certain frustration when performing the task.

8.1.3. Machine Translation

During the interviews, the translators made several references to machine translation. Some of them had had positive experiences, others had not. The majority had mixed feelings about machine translation. The term here “mixed” means that translators thought that on some occasions the experience had been good and on others poor; it does not mean that they were unsure or had doubts about the MT output quality in a given project. Three translators made reference to their positive experiences because, although in some cases, especially with long sentences, the task was complex, if MT output was used correctly it could be “very dynamic”, “it is faster”, “less monotonous”, “interesting”. Eleven reported having had mixed experiences in the past. For example, they found that on some occasions the terminology was perfect, but the sentence structure was very poor, that sometimes the whole sentence had to be “reshuffled” but in other cases the result speeded up their work, or that in some cases the results were “terrible” but in others, as in this project, the results were good; that sometimes “you don’t have to intervene really but other times you go crazy”, or that sometimes MT is better, sometimes it is worse, so that MT becomes difficult to quantify.

Translator 15 made an interesting comment: “In general machine translation does not need to be perfect but only understandable”. In his view, MT was more beneficial to him financially if the quality requested was “understandable” and where he did not

“have to worry too much about the style”, that is, if customers used MT for material that was not highly visible. On the other hand, if the quality expected was very high, then, he felt he had to make many changes (style and terminology) and it became unprofitable. In other words, it seems that Translator 15 preferred to use MT for “fast post-editing” rather than for “full post-editing” (Allen 2003), and that customers should be more flexible in their style and terminology requests if they are using MT in their localization process.

Translator 20 made a relevant remark regarding tags in the documents: “You work with the variable tags and you always have to touch the segments, you always have to change the order of something”. The fact that this text was free of tags could certainly be a factor that would speed up the productivity process for translators in both Fuzzy and MT matches. In our experience in the commercial sector, translators often complain that with a heavily tagged document it is easier to work from the source text and not from a proposed text where tags need to be rearranged completely in each segment. Translator 20 also commented in the questionnaire, “I find the process unfairly paid at times, and I miss the creative feeling, even if it is software manuals, that translating from scratch brings.”

Translator 23 summarized well the feelings that some translators expressed during the debriefings, particularly in relation to the varying quality of MT. She also expressed an interesting opinion on this type of study:

[...] a lot of importance is placed on time employed in doing the job and I think this sometimes goes against the translator because there are sentences that are easier than others, or depending on the translator’s experience he would go faster or slower [...] I think that machine translation should be considered from the linguistic point of view almost exclusively.

Although doing research or simply measuring the use of MT from a time point of view is insufficient, we cannot negate the fact that the use of MT is directly related to speed (reaching markets more quickly), volume (more content in more languages) and saving costs. We agree that time on its own only tells part of the story (as we have clearly seen in the quantitative analysis) but time is nevertheless an essential part of this story. As Translator 23 remarks, analyzing MT is a very complex topic and many factors are involved such as quality of the output, experience, training, purpose of the post-editing job or even quality of the “translator”.

There were four translators that gave clearly negative feedback about MT (Translators 14, 17, 20 and 22). They had also responded in the questionnaire that they were Indifferent, Dislike or Strongly Dislike using MT as part of the localization process. There were several reasons: “projects are full of instructions and a lot of glossaries to follow”, “technical aspects that present obstacles”, proposals are “so bad” that the translator had to return the assignment, the rate is lower than the effort required, or segments have to be completely redone. Translator 20 openly said that she was not a “fan of machine translation”. Her reasons are interesting:

I thought I was working for little money for the time I had to invest in that type of translation and then also it is a personal preference because I don't like revising in general. I prefer a process of creating from zero, to translate.... There might be people that prefer to revise, I don't know... There was this customer that I'm thinking of right now that pays us the same rate as *Trados fuzzies* [English in original] but really I can't tell you if the effort is equal, lower, higher but I have the impression that I have to stop more and I don't trust it as much.

It is interesting to see that Translator 20 does not like revising and prefers to create something from zero, and this is regardless of her productivity when doing so or the rate. Also, she does not know but she feels her effort is higher when post-editing than when editing TM fuzzy matches, although there is evidence (as we have reported in this study) that MT correlates well with TM fuzzy matches in terms of “time” effort. We could hypothesize that, for translators, if the cognitive effort is higher with certain MT segments (Krings 2001, O'Brien 2006b), their perception of the whole post-editing exercise is that it takes longer, although this might not actually be the case in terms of temporal effort. Finally, Translator 20 stated that she trusted a Fuzzy match segments because it comes from a human translation but did not trust something that came from a machine. We have already noted that some translators might trust fuzzy match segments more readily (Guerberof 2008) and also that the provenance information although it does not affect the overall speed of an assignment or the quality, might be relevant for translators working on individual segment types (Teixeira 2011).

In general, translators show some knowledge of how machine translation works and its error typology. For example, “some of the errors had to do with problems with structure, order, that tend to be typical of machine translation”; “MT is useful in some cases and not in others”. In short, from the debriefings it was obvious that translators

had experience of this task but also that they had a professional outlook on the topic rather than an emotional one. For example, Translator 14 presented an interesting view of the current situation for translators and MT:

I think that apart from the fact that you might like it or not, that you feel comfortable or not, these are the trends in the current market and we have to get familiar and up to date because it is what it is being used at this moment. So, many times, one prefers other types of work but if you are not up to date and learn new tools and up to date with machine translation, the current market now, you are left out. This is, I think, a reality.

In conclusion, the group was highly familiar with machine translation and their attitudes were open and flexible. This does not mean, however, that they liked using MT. This signals a change with respect to previous views on how translators perceive MT (Schäler 1998, He et al. 2010a, 2010b, Carl et al. 2011, and many others) where translators are seen as very reluctant to adopt MT as a working tool. In a 2010 survey about Post-editing, TAUS mentions translators' resistance as one of the main pains in post-editing management. Later, in their post-editing report (TAUS 2010b), they try to explain this resistance by suggesting that post-editing requires a higher cognitive load than translation and therefore it would be understandable for translators to show some kind of resistance. They also explain that for translators, dealing with MT is "similar to the emergence of TM tools in 1990" (TAUS 2010b: 15) and would be like dealing with a poor TM tool. Evidently, the opinions in the survey are from companies engaging in this type of activity and how they perceive translators' attitudes; this was not a survey designed to gather information from translators. The results might have been completely different if translators had been asked. Also, we are not entirely sure - and translators in this project do not seem to be either - that dealing with poor TMs is the same as post-editing MT. In our project, the translators seem to have had a very practical and open attitude towards MT, although some did not like working with it for different reasons. Tatsumi (2010: 185) has already commented on this in her thesis: "the answers to our questionnaire suggest that a flexible and down-to-earth attitude towards PE is the trend", and this was also the case in this study. Also, Lagoudaki (2008) in a survey conducted about the value of MT for the professional translators concludes, "machine translation appeared to be well received amongst translators who were familiar with it" (ibid: 265) and also "translators also seem to be coming to terms with machine translation as an alternative means of translation production" (ibid: 268).

8.2. Reviewers' debriefings

We were interested in knowing the reviewers' opinions of the assignment and how they approached the task. Given that they had 24 translations of the same source text to review, we wanted to know if they had picked up any anomalies during the process or if something had caught their attention. Further, since the reviewers are also translators and post-editors, we were interested in knowing their opinions about the assignment and if they had something to add to what had been already stated during the questionnaire.

The questions we asked the reviewers were:

- What did you think of the instructions for the task including the glossary?
- How did you find the review methodology proposed?
- How do you normally review translations?
- Which errors were difficult to classify?
- Of the 24 translations, were there any that caught your attention for a particular reason?
- Did you find that the translations were similar with respect to the type of errors made?
- Did you find a lot of over corrections?
- Was there any segment that was difficult to translate?
- Which questions in the questionnaire were difficult to answer?
- How did you feel when completing the task?
- Would you like to add anything else?

The full translated and transcribed versions of the interviews can be found in Appendix J. As with the translators, during the debriefings some topics emerged apart from the ones we were initially proposing during the semi-structured interview. The final framework that we setup with NVivo was:

1. Assignment
 - 1.1. Instructions & Methodology
 - 1.2. Glossary
 - 1.3. Questionnaire
 - 1.4. Translations
2. Feelings
3. Revision process

8.2.1. Assignment

8.2.1.1. Instructions and methodology

The reviewers found the instructions clear and the methodology easy to follow. Reviewer 2 mentioned that she liked the fact that she could use Track Changes during the review because this option allowed her to go over her corrections, see what she had done before and modify if necessary. However, Reviewer 2 had problems initially with the time tracker since every time she spotted an error she stopped to include the time in the form. She realized, however, that this was time-consuming and she decided to finish correcting the Word document per translator before transferring the error result to the LISA QA form. Reviewer 3 thought that some information was missing regarding linguistic aspects such as the use of the infinitive and the steps to take with descriptive translations and software options. She found it was difficult to have clear criteria for these two aspects. Also, she found it hard to work without a translation memory.

The reviewers nevertheless mentioned passages in the instructions that were unclear for translators: the use of the imperative, infinitive and gerund; the use of upper and lower cases (although Reviewer 1 mentioned that this for a couple of translators only and Reviewer 2 mentioned that it was very clearly explained in the glossary), pressure not to make unnecessary changes and at the same time seek for publishable quality, and issues with software option translations (Reviewer 3 was unsure if descriptive translation of software options should be marked as correct translations, even if the glossary contain the exact software option translated). And finally, it emerged during the debriefings that the instructions were unclear about the possibility of making queries that are normally directed to customers in a commercial context.

8.2.1.2. Glossary

The three reviewers mentioned that the glossary was very useful and complete. Reviewer 2 was even surprised that a translator had made mistakes in upper and lower cases when the glossary contained all options.

However, the glossary contained words that translators would not normally search for because they seem fairly standard. Reviewer 1 gave a clear example: email as *mensaje electrónico* instead of *correo electrónico*. The majority of translators had it wrong precisely because they did not look for it (as seen above, Translator 12 gave the example of *añadir* and *agregar* for “to add”). The reviewers thus remarked that the glossary might have been too detailed.

8.2.1.3. *Questionnaire*

The reviewers found the questionnaire interesting and easy to complete. Reviewer 3 mentioned, “I liked the survey; I thought it was very interesting because we don’t stop to analyze [the processes]”. However, they were surprised at those questions that they found difficult to answer. For example, Reviewer 2 commented:

I had never stopped to think [about this]. Because the rates are not very negotiable, you don’t stop to think [...] Am I satisfied?

Perhaps some translators ask themselves if their work is profitable globally (at the end of a working month, for example), but they do not necessarily ask if each specific rate per word is “satisfying”. Notwithstanding this, it is peculiar that the rate is not the main concern for Reviewer 2: in the questionnaire she responded that she was “Satisfied” with the rate and “Very satisfied” with the work as a reviewer, so it is somewhat natural that price is not her main preoccupation. Reviewers 1 and 3 had problems answering questions about methodology. Reviewer 1 found it difficult to offer a response when different alternatives were given because one option could imply another option: “For example, Read the source and then the target, and then, Read the source, the memory and the target”. Reviewer 3 had doubts as to her exact review method. She was unsure whether she read the source first or the target. She had to stop and think how she manages the text to review.

8.2.1.4. *Translations*

Reviewers commented on several difficulties that translators had to face when working on the assignment. It was difficult to determine when to use the imperative and when to use the infinitive. Reviewer 1 decided that if the sentence ended with a period then the imperative was more appropriate. This is somewhat strange since the instructions were to use the infinitive when in doubt (see Appendix C). This is in line with findings reported from Depraetere (2010) and Belam (2003), where translation students were uncomfortable with grey areas and with unclear instructions. In this case, we can see that Reviewer 1 had to create his own rule or instruction, despite having more flexibility. Reviewer 1 also commented that those translators that accepted MT proposals without major post-editing, tended to either change the meaning or use incorrect terminology, while in the case of Fuzzy matches, they were only required to change one word and fewer errors were made. This could certainly be the case for

certain segments. If we check the absolute number of errors (see section 5.4) some translators showed more errors in the MT category than in the Fuzzy match (in 15 cases). Although statistically significant differences were not found in these two categories considering the text length, Reviewer 1 did perceive the overall number of errors being higher in MT for some translators.

The reviewers also commented on the difficulties that the lack of context presented for translators and reviewers alike. Reviewer 1 commented that it was part of the revision process. Lack of context is still an issue not only in this project but more generally in the type of text that translators deal with in the localization industry.

Another related issue that emerged from our interviews was that the translators and reviewers did not ask queries about the text, because they were under the impression that it was not possible to do so. The email sent at the beginning of the project specified that translators could ask us questions throughout the entire project. Perhaps, the fact that this was not specified as such in the instructions led them to believe that they could not contact us for this type of linguistic query but only for procedural questions (duration, technical aspects and fees, for example).

There were also certain segments that reviewers thought were conflictive because the source text was ambiguous. These were segments 78 and 118. Segment 78 (“the display style and color of the gauge faceplate, gauge border and needle”) presented two problems according to reviewers. For Reviewer 1, it was difficult to know if the “needle” referred to the “gauge” or not, and for Reviewer 3 it was unclear if “color” was also referring to “display”. Segment 118 (“Any kind of selector except metric condition selectors, which filter metric values and ranks, can filter or slice”) presented problems for Reviewer 2, as she was unsure if the “which” was a referring to “any kind of selected” or “metric condition selectors”. These two source texts were translated using MT. As we can see here, issues in the source text might affect the quality of the MT output (as has been observed by O’Brien 2006a, Tatsumi 2009, and Tatsumi and Roturier 2010). Although the source text was already ambiguous and reviewers had to read over the sentence several times to understand it, the MT output might have contributed to the confusion in rendering the wrong word order.

The type of error that Reviewer 1 reported as being most problematic for translators was the use of the *gerundio* (Spanish gerund) to express consequence rather than simultaneity. We discussed this in section 5.2.3. Reviewer 2 commented on this same topic:

The issue with the *gerundio*, for example, I think that in my case, it was a reviewer that I had many years ago that she made me “*anti-gerundio*” but sometimes they are correct, and I have to continue to consult when they are correct and when they are not.

Regarding error classification, Reviewer 1 mentioned problems classifying errors as Accuracy or Style errors, or between Language and Style, and in general he pointed out that the one error “refers to more than one classification. So sometimes it is difficult to evaluate them and see which one is more important”. Reviewer 2 had more problems between Mistranslation and Accuracy because it was not clear to her if the translator had not “fine-tuned well or if he had not understood correctly”. She also mentioned that, on certain occasions, more than one translation could be valid (especially with the lack of context) and that she would expect translators to be consistent in one single “version” of the English source. However, she did not consider these inconsistencies to be errors if all versions were valid. Reviewer 3 mentioned that she did not have difficulties although sometimes Mistranslations tended to be more difficult to classify.

The translators that performed well were 12, 20 and 22 according to Reviewers 1 and 3, as we saw in the questionnaire. Reviewer 3 mentioned that there were only two or three translations that were “good”. On the other hand, Translators 10, 18 and 19 had a high number of terminology errors. Reviewer 2 also mentioned that certain translators showed a lack of experience in this domain. Translator 19 is a relatively novice translator and Translator 10 has between two and four years’ of experience. However, Translator 18 has between six and eight years’ of experience, but as we saw above, the instructions were not altogether clear to this translator and he declared that he did not translate in the questionnaire (he post-edited and revised translations).

The three reviewers agreed that there were not many overcorrections: if changes were made, they were justified.

8.2.2. *Feelings*

The three reviewers felt comfortable doing the task. Reviewer 1, although he commented that “after doing 15, 16 or 20 times, one wants to finish”, mentioned that “you can tell that you are maturing a little bit the translation”, referring to the fact that as he progressed he discovered that new translations helped him clarify meanings and he had to go back and change his previous corrections. Reviewer 2 felt very good and

liked the fact that with Track Changes she could check the entire file before closing it. Like Reviewer 1, she also remarked on the learning process during the revision:

Considering it was the same text, there were quite a lot of differences. That surprised me, how we can interpret the same sentence in so many different ways. [...] When you look at a sentence, you give it a meaning for your context, your experience, for whatever reason, you give it a meaning, but if you look at it 7, 8 or 24 times you realize that really it could have another meaning.

This clearly signals the complexity of reviewing in the localization context. Reviewer 3 felt comfortable because, regardless of the volume of work being high (48,000 words to review), the text was repeated and the forms were easy to use.

Reviewer 2 mentioned that in the beginning she felt some pressure and she had the impression that she was not following the instructions correctly, but after a few files this “feeling” disappeared. Reviewer 3 commented that sometimes classifying the error in the three categories (Fuzzy, MT and No match) could be cumbersome but not difficult.

8.2.3. *Review process*

During the debriefings, the reviewers explained how they review translations. Reviewer 1 reads the source text first, then the translation memory, the target text and finally he checks terminology. If he is doing a final revision, he only concentrates on the final target text and he only consults the source when he has doubts or something “sounds unnatural”. He mentions that being familiar with a particular customers’ terminology or even vocabulary related to a particular product helps the revision as well as the translation process, and the resulting errors are fewer.

Reviewer 2, however, reads the target text first, compares it to the source and then makes the changes. But if she has a complete document she would read the original text to see what it is about and to check the style, and then go to the target text. She commented that with tools such as SDLX where there is no context and she needs to revise single sentences, “you are a bit lost, so I prefer to work in context”, so she asks for the original files in order to check this context. On the topic of the review forms, Reviewer 2 mentioned that it takes quite a lot of time to fill them in, and she suggested that having Track Changes or a feature similar to the ones in the tools that she normally works with would facilitate the review process.

Reviewer 3 also compares the target text to the source text and if there are several related files, she works to the end of the batch and then she starts again to correct issues she might have clarified along the way. She checks terminology and doubts as she progresses but if a particular issue is difficult or very time-consuming she takes notes in a list of queries and continues, then she clarifies the complete list of queries at the end of the batch (perhaps other files have the answer to that particular doubt). Reviewer 3 also commented that she does not tend to trust translation memories because they might contain different product versions, for example, and terminology might change from one version to the next, so she would rather work with the latest memory belonging to that version of the product rather than work with a global one.

The reviewers all agree that their revision process depends heavily on the type of project they are working on as well as the technology, and that they would adapt depending on this.

8.3. Conclusion on qualitative results

The translators found the instructions easy to follow, appropriately short and to the point. There were, however, issues that emerged during the debriefings such as the use of the imperative or the infinitive in the translation of the source, the use of upper or lower cases in the translation of software options, and the type of changes allowed in order to attain publishable quality while at the same time not introducing preferential changes. This might have resulted in quality problems for certain translators (for example, Translator 18). They found the glossary easy to use and surprisingly complete. However, on occasions the glossary was so detailed that translators failed to look for the term, or the standard translation was not the one used in the glossary. The fact that the glossary was in Excel and not integrated into the tool delayed the task. This could be a partial explanation for the high number of terminology errors. Overall, they were comfortable with the assignment, as it was short and straight forward, although some translators were uncomfortable because of the methodology imposed or the tool used. This was the case for Translator 13, and this could explain the high number of errors this translator made. The translators did not encounter major problems when connecting or manipulating the tool. Some of them found the tool attractive because it was easy to learn and it presented a simple interface, while they did miss certain features they are accustomed to such as a spell-checker, being able to revise segments once they are

closed, having the words changed highlighted, having the glossary integrated into the tool, and having the source text pasted in the target text by default. The tool did share and lack characteristics with commercial tools at the same time. The text did not present major difficulties and most translators found the proposals (MT and Fuzzy matches) useful. They could identify different characteristics in these two types of texts, despite not knowing which segments were Fuzzy match and which ones were MT output. We did notice, however, that translators tended to identify “good” segments with human translations (as in He et al. 2010b), though some MT segments were left unchanged during post-editing. There were segments that were difficult to translate due to the lack of context or because of ambiguity in the source text, but this did not constitute the norm. The questionnaire was also easy to understand and complete, but there were certain difficulties, mainly questions about rates because of all the variables involved in payment, about revision methods because of the variability according to projects and in some cases the unawareness of their own work methodology, and about post-editing because of the variables involved but also because of the lack of knowledge of the subject (for those translators with no experience in post-editing).

All translators except one were satisfied with the work they do, but not necessarily with the payment they receive for different tasks, although this was highly dependent on different customers. The payment for Fuzzy and MT matches might be inadequate if the quality of the translation memory or MT output is poor and the translator has to invest more time in fixing those segments than if they did the translation from scratch, while they are paid only a fraction of the word rate for translation. It was not clear if they prefer payment per hour or per word, but translators did indicate the need for a payment related to the quality of the MT output or TM, or to the nature of the task requested. The methodology for reviewing (texts and fuzzy matches) tends to be to open the segment, read the source, apply changes to the target and check the tool to see the changes marked. Before handing back the files, these translators would recheck the batch of files received. Reviewing after a days’ work or after finishing a file is less frequent. There were several problematic issues that they signaled in the translation process: the excessive number of instructions to complete small tasks, terminology maintenance, and excessive reference material, and tagging in documents that force translators to rearrange every single segment regardless of the level of fuzzy match or quality of the MT output.

This group of translators was in general quite familiar with machine translation and post-editing, but not all of them were performing these tasks on a regular basis. They could identify clear MT segments and knew what to change in those cases. Although some did not like doing post-editing, mainly because the quality of certain MT segments was poor or the instructions too cumbersome to follow, or they did not like to review, the overall attitude was nevertheless flexible and practical. The translators that dislike post-editing would in general not perform the task, and those that post-edit find that experience helps them spot errors and that in some cases it increases their productivity. Post-editors do not feel that they grow accustomed to MT errors or that their productivity decreases over time. Most find that post-editing requires either similar or more effort than editing human translation, and this could refer not just to a higher cognitive effort for this task (not necessarily a temporal effort), but also to the fact that each translator might have different experiences with previous post-editing jobs and might also perform differently because of their own personal characteristics. Also, many were aware that post-editing will be a necessary task in the future of localization and that outputs will improve over time. From this group of professional translators we can see that those doing post-editing are well-informed about the process and the current shortcomings. We do not find a negative attitude towards working with MT (although the majority of translators might dislike it) but rather problems with how the task is paid or organized.

Reviewers also found the instructions and methodology easy to use. There were several issues, however, with the imperative and the infinitive, the use of upper and lower case for software options, and descriptive translation of software options. The use of Track Changes was comfortable and useful for the reviewers as they could see the changes made, but not having a translation memory was a problem for Reviewer 3 as she could not consult previous decisions quickly. The glossary was very complete although the reviewers remarked that some terms had not been consulted by translators, and others did not follow the glossary at all. The questionnaire was also easy to complete, although the reviewers had problems answering questions about methodology and rates, mainly because the answers depended on several factors. The reviewers thought that the general quality of the translations was “Average” and this is easily explained by the differences of error among translators, some delivering good quality (Translators 12, 20 and 22) while others delivered poorer quality (Translators 10, 18 and 19). The reviewers thought that some segments were difficult to translate (for example

segments 78 and 118) because the source text was ambiguous and this caused difficulties for translators. They were also surprised at the number of terminological errors, incorrect translations of prepositions and the use of the *gerundio* (although there were different opinions on the latter). The reviewers thought that the LISA form was easy to use and that it does not require changes. However, it emerged that it was difficult to separate between Mistranslation and Accuracy, one error can be classified under several categories making it difficult to decide where to place it, and classifying errors according the different Match categories was also cumbersome (in this particular project). The reviewers felt comfortable doing the task, although there was certain pressure at the beginning until they became familiar with the procedures.

The reviewers, like the translators, were not satisfied in general with the payment they received for the task of reviewing, and they thought that for Fuzzy and MT matches it depended on many factors but it was sometimes unfair. They either dislike or were indifferent to working with MT, mainly because it was highly dependent on the quality of the MT output, the tagging in the text, and the quality of the terminology proposed. The reviewers said they had different revision methods, although they all thoroughly check each term against the glossary.

PART IV: Conclusions

This part presents the conclusions drawn from both the quantitative and the qualitative results in relation to our research questions and the related hypotheses.

Chapter 9: Final conclusions

This chapter presents the conclusions from the quantitative and qualitative analysis, as well as lines for possible future research.

9.1. Conclusions of combined results

We set out to find support for three hypotheses and three sub-hypotheses that would, in turn, answer our three research questions. After conducting this experiment and analyzing the data, we can review our final findings. Let us examine first, the hypothesis on productivity:

The time invested in post-editing machine-translated text will correspond to the time invested in editing fuzzy-matched text corresponding to the 85-94 percent range.

The data support this hypothesis. The time invested in processing MT match segments is not statistically different from the time invested in processing Fuzzy match segments in the 85-94 percent range. Further, the results show statistical differences in speed between No match, on the one hand, and Fuzzy and MT matches on the other, indicating that the texts proposed help translators increase their productivity.

We also sought to investigate adequate levels of payment for MT matches, and several conclusions can be drawn from this experiment. If the engine is trained with sufficient (high number of bilingual or parallel data) and cleaned translation memories and the automatic and/or human evaluation scores are high (in our case the BLEU score was 0.6 and the human evaluation 4.5 out of 5), translators show similar productivity in processing MT output as in processing 85-94 Fuzzy matches. The productivity gain and associated time savings vary considerably for each translator and this indicates that some translators benefit more than others from these translation proposals. The average time savings for these translators (32 percent for Fuzzy matches and 37 percent for MT

matches) is lower than the average 40 percent assumed for this type of matches (85-94 fuzzy matches) in the industry but this might be explained in some cases by the use of a different tool that does not highlight the changes in the Fuzzy match segments. We also found out during our analysis of the qualitative data that translators might not always be exposed to high quality MT output, and perhaps because of this, they think the effort required to post-edit MT segments is higher than that required to edit human translations. This could also mean that the perceived cognitive effort for MT is higher, and therefore, translators think that the temporal effort is also higher. Consequently, they are not fully satisfied with the pricing scheme in the industry. Why is this? Their level of satisfaction seems to be higher with TMs than with MT output (although the difference is not pronounced) precisely because the quality of the MT outputs varies significantly from project to project, or even from segment to segment. Price is a factor that worries this group of translators, which is otherwise quite satisfied with their profession. We cannot establish if a price per word or per hour is preferred, but mostly that the price should be in line with the quality of the MT output or TM used. This might appear to be an obvious conclusion, and it is common to hear in MT-related discussions that “garbage in” means “garbage out”, referring to the fact that if the quality of the material used to train the engine is low, the output will consequently be low. In this experiment we have seen to what extent this is true for productivity and quality. Another issue that affects their productivity is excessive tagging, excess of reference material, wrong terminology, discrepancy between quality and productivity expected. Our informants also suggested that when approved terminology differs from the terminology found in MT output or TMs, and instructions are too cumbersome to follow and not sufficiently clear, their productivity is negatively affected. It seems that either terminology is managed adequately from project to project, or decisions are made to be more flexible about quality to ease the task for translators, especially if discounts are expected. We can conclude that the productivity of MT matches can be like that of 85-94 percent TM matches if the MT output is of high quality, tagging is not excessive, instructions are short and clear, terminology is either final in the output or few glossaries need to be consulted (as in this experiment), and the quality expected is in line with the quality of the output.

We also formulated a sub-hypothesis:

The translators with higher processing speeds, in words per minute, when translating the “No match” segments will have less productivity gain when post-editing the proposed text from MT or TM than the translators with lower processing speeds when working with the same set of segments.

The data do not support this sub-hypothesis. The translators were divided into different Speed groups (according to their No match average processing speeds) and, although the productivity gain slightly decreased in the fastest group, there were no statistically significant differences among the three Speed groups. The productivity gain obtained with Fuzzy and MT matches does not seem to be necessarily related to the intrinsic speed (No match speed). However, there is high inter-subject variance and this would imply that some translators might not be satisfied with the discounts they are asked to offer because this discount might be proportionally greater than the productivity gain they obtain with the TM or MT proposals. There is no clear solution for this dilemma, unless times and quality are measured for each translator after each project, and this seems an unfeasible alternative. However, if a translator declares that a particular TM or MT output does not improve his or her productivity, this might actually be the case for this particular translator. We also found out that if the intrinsic translators' speed is considered (No match), there are statistically significant differences between the MT match and Fuzzy match and that the processing speed for MT is significantly higher.

TER values were also examined to establish how many edits translators made to both Fuzzy and MT matches. There were statistically significant differences between the two types of matches, indicating that the translators had made significantly fewer edits when processing MT matches. Despite having a high quality output on average, translators still changed 80 percent of all strings. In the qualitative analysis, we learnt that some translators would assume that “good” translations were always “human” translations despite the TER value indicating that 20 percent of MT-match segments were not changed at all. We also examine correlations between TER and time and we further clarify that the number of edits at lower and medium speeds seems to change from translator to translator, and not so at high speeds, so this might indicate that specific instructions should be given on how to deal with segments that require a significant number of changes. The TER analysis signals a possible solution in pricing, globally rather than at segment level: if TER is applied to material that is already post-

edited, and there is a high level of MT matches unchanged or a low TER score (as we have seen in this project), this would help to calculate a possible percentage discount to apply in future projects if the conditions are of similar nature. However, this correlation between time and TER is not always applicable in individual segments, some segments with a low TER (fewer edits) had a low processing speed and vice versa.

Let us turn to the hypothesis and sub-hypothesis related to quality.

The final quality of the revised target segments translated using MT technology is higher, if measured in number of errors, than the final quality of revised Fuzzy match segments and lower than the final quality of revised No match segments.

The data do not support this hypothesis. The number of aggregated errors from three reviewers show that the final quality of the post-edited MT segments is not statistically different from the final quality of the edited Fuzzy match segments and it is significantly higher than the translated No match segments. Moreover, we observe that there are more Language, Terminology and Style errors in the No match segments, more Accuracy errors in Fuzzy matches and more Mistranslation errors in MT matches. Further, Fuzzy match segments show fewer Style errors and MT match segments show fewer Terminology errors than the other two categories. These results might be surprising because human translation quality (in this case the No match) might be expected to give higher quality than post-edited translation (MT match). However, it is really quite logical judging from the quality of the original material (translation memories) used to train the engine and to create the Fuzzy matches, and given that the No match segments were only translated and not revised by a third party. It is also interesting to see that the reviewers changed 20 percent of all strings as opposed to the 80 percent that the translators edited.

We were trying to find out if MT had an impact on the final quality of the post-edited text. We can conclude that in this experiment both the MT and TM proposals had a positive impact on the quality since the translators had significantly more errors in the No match category, translating on their own, than in the MT and Fuzzy match categories. The qualitative analysis showed us that the high quality of the MT output was influential in the final errors obtained in these categories. It also shows that there are certain factors that might have influenced the translators' quality negatively: the fact that they could not go back to translated or post-edited segments, that they did not have

a context for the segments, that the glossary was not integrated into the tool, that the source text contained ambiguous structures, and that the instructions might have been too vague for certain translators. These factors highlight several issues to consider when measuring quality, and when organizing projects.

The qualitative data shows that translators might be asked to produce top quality when the starting point for the project is of very low quality, and this causes frustration and price dissatisfaction. The qualitative data also shows that translators might believe they have been thorough in performing a certain task (checking the glossary, for example) but this does not correspond to the actual results obtained from reviewers.

The revision phase showed a great disparity in reviewers' corrections. The reviewers did agree on the No match category, but the agreement on Fuzzy and MT matches was either weak or there was no agreement, perhaps indicating that the origin of the text might have influenced their evaluation. The reviewers also tended to agree on best and worst performers in general, but there was great disparity in the translators' classifications if they were ranked according to the number of errors. These disagreements could also mean that each reviewer might adapt the instructions to their own particular logic if grey areas are perceived, or that each reviewer focuses on areas of particular interest (for example, certain grammatical errors) that they have established from their previous experience, or that the source text can be interpreted in different ways, especially in the absence of context. A more in-depth analysis of the corrections is needed to further explore these findings.

Our Quality sub-hypothesis said that:

Translators with overall higher processing speeds, when using MT or TM technology, will have fewer errors than those with lower processing speeds.

The data do not support this sub-hypothesis. The translators were divided into different Speed groups according to their No match processing speed and also according to their global processing speed and there were no statistically significant differences between the groups as far as errors were concerned. We observed that there were fast translators with fewer errors and slow translators with many errors, but also fast translators with many errors, and slow translators with fewer errors. Therefore it is not clear that spending more time on a translation might give better quality results, although this could be the case for certain translators. The final quality seems to be related to the

translation proposal, the translators' skills, and also to the reviewer's particular revising style. Quality depends not only on the translator, as we have seen, but also on the reviewer.

Let us look now at the hypothesis related to experience:

The greater the experience of the translator, the greater the productivity in post-editing MT match and Fuzzy match segments.

The results do not support this hypothesis. The translators were divided into four different clusters according to their experience as determined from their responses to an on-line questionnaire. The results show no statistically significant differences between the clusters as far as speed is concerned. The groups with greater experience show higher productivity than the ones with medium experience; the group with the least experience, however, shows the greatest productivity. We also found that the cluster with no or little experience in post-editing was slower in processing all three matches. The low processing speed in MT matches could be explained by the fact that they did not have post-editing experience. The low processing speed in No match, and possibly in the other two categories, could be explained by the fact that this group contained translators with the lowest reported typing speed.

The sub-hypothesis on experience said that:

This experience will not have an impact on the quality (measured in number of errors).

The results do not support this sub-hypothesis. The cluster with the least experience has more errors than the other three clusters, and this difference is statistically significant in Fuzzy match and No match. Despite the fact that the errors in MT match are higher in this cluster, the difference is not statistically significant, showing that MT had a "leveling" effect with more novice translators. This could imply that a high quality MT output could be used instead of Fuzzy matches in the 84-95 percent range for translators with less experience in order to produce better quality results. The data indicate that the novice group was not as thorough as the other three groups in following both the instructions and the glossary provided. The cluster with the most experience in post-editing shows fewer errors in the MT match category,

indicating that experience do have an impact on quality. The group with little or no post-editing experience had a low number of errors in Fuzzy match, suggesting they were slower because they were also more thorough.

We were investigating whether experience was influential in the post-editing of MT outputs and our data shows that experience can be influential in terms of quality: the senior translators appear to be more thorough when following instructions (especially glossaries) than junior translators. We observed that novice translators were faster, although not significantly, when processing the segments, although the number of errors left remaining was significantly higher than the number of errors found in the work done by more senior translators: correcting errors and achieving a final quality according to the instructions requires certain experience. The groups with more experience in post-editing performed faster than those without it, but they did not necessarily have fewer errors. The qualitative data also tell us that the translators thought that experience did not necessarily increase their productivity when post-editing, although it did for some of them. However, they did find that experience helped them to spot errors, and that exposure to post-editing did not affect their ability to correct MT errors. Therefore, we could conclude that the ideal profile of a post-editor is a translator with more than two years experience in localization and also with experience and training in post-editing. If the post-editors are less experienced, emphasis should be placed on following the instructions and terminology provided, and being thorough in the editing of the proposed segments.

Finally, during our qualitative analysis we found that the translators, including the three reviewers, had ample knowledge about MT and post-editing despite the fact that the majority of them did not like working with it or had mixed feelings about it primarily due to exposure to poor MT output or project set-up as explained above. Further, some translators simply did not like reviewing in general and they preferred to translate from “scratch”. They were aware, however, that post-editing was the future in the localization industry and that they would need to adapt to the new market needs. This might indicate a change in translators’ opinion and acceptance of MT output as part of the localization workflow.

9.2. Further research

This research project opens up many different lines for further research. Here are some we could envisage.

Perhaps the most obvious extension could be to use this same methodology with several language combinations to see if the results are similar or to what extent they are language dependent. Similarly, the method could be used with different content or with engines giving different BLEU scores, in order to explore how scores related to final quality.

With the data obtained from our TER analysis, we could examine the type and number of edits that the 24 translators made in the MT and TM output (in terms of Deletions, Shifts, Insertion, and Substitutions) with a view to exploring further if changes made by translators are similar, essential or preferential.

A study could be set-up to explore in more depth the value added of traditional revision cycle in localization. If, as we have seen, reviewers essentially disagree, then this raises a question about the process. In particular, if more companies move towards MT and post-editing, as is likely, a question emerges over the need for the traditional review cycle. With the data gathered in this project, we can analyze the corrections made by the three reviewers and find the levels of agreement and disagreement in terms of number of errors per segment and per translator as well as the classification of these errors.

Our methodology could be combined with an eye-tracking tool and key-logging with the aim of measuring the cognitive and temporal effort when processing TM and MT segments, and to explore the relationship between how translators perceive temporal and cognitive effort and actual temporal and cognitive effort.

Another valuable perspective would be to examine the views that end-users have on material produced in this new “hybrid” environment, especially in comparison to the views of professional translators. We could analyze how users rate final target texts resulting from post-editing in comparison to how professional translators and reviewers rate these target texts.

A similar study could be done using two groups of participants: senior and novice professionals. We could observe the differences in post-editing strategies and resulting quality and productivity, using this present methodology and retrospective interviews. We could then describe the productivity and common errors made by novice and by

professionals when post-editing, thus potentially improving both professionals' practice and novices' training. In the same line of research, a comparison between amateur and professional translators could be made.

Another study could be designed to analyze those translators that have high processing speeds and low number of errors. A group of translators could be selected and tested according to their speed and number of errors. Then translators from this initial group can be selected to form two groups: the ones that have a high processing speed and low number of errors, and those that have a similar processing speed but higher number of errors. Eye-tracking and key-logging can be used with both groups while working and results can be contrasted, partly to see if they are using different translation strategies.

Finally, a thorough qualitative research could be carried out (by means of a survey or interviews) of translators' views on emerging translation practices. This might focus on CAT tools/MT, instructions, reference material, and pricing, as well as on translators' perceived difficulties, opinions and suggestions for improvement.

References

- Allen, J. 2001. "Post-editing: an integrated part of a translation software program". *Language International*. 13 (2): 26-29. Available from <http://www.oocities.org/mtpostediting/Allen-LI-article-Reverso.pdf>. Accessed June 2012.
- Allen, J. 2003. "Post-editing". In *Computers and Translation: A Translator's Guide*. Harold Somers, ed. Amsterdam and Philadelphia: Benjamins. 297-317.
- Allen, J. 2004a. "Case study: implementing MT for the translation of pre-sales marketing and post-sales software deployment documentation at Mycom International". *Machine Translation: From real users to research. 6th Conference of the Association for MT in the Americas*. Frederking, R. Taylor, K, eds. Verlag. Berlin. Heidelberg: Springer. 1-6.
- Allen, J. 2004b. "Perspectives on Machine Translation". *The Guide from Multilingual Computing & Technology*. 62: 8-10.
- Allen, J. 2005a. "An introduction to using MT software". *The Guide from Multilingual Computing & Technology*. 69: 8-12
- Allen, J. 2005b. "What is post-editing?" *Translation Automation*. 4: 1-5.
- Alves, F. Liparini Campos, T. 2009. "Translation technology in time: investigating the impact of translation memory systems and time pressure on types of internal and external support". Gopferich, S. Jakobsen, A. Mees, I, eds. *In Behind the mind. Methods, models and results in translation process research*. (Copenhagen Studies in Language 37). Copenhagen: Samfundslitteratur. 191-218.
- Andrés-Lange, C. Scott Bennett, W. 2000. "Combining Machine Translation with Translation Memory at Baan". In *Translating Into Success. Cutting edge strategies for going multilingual in a global age*. Robert C. Sprung, ed. Amsterdam and Philadelphia: Benjamins. 203-218
- Arnold, D. Balkan, L. Meijer, S. Humphreys, R. Sadler, L. 1994. *Machine Translation. An Introductory Guide*. London: NCC Blackwell. Also available from <http://www.essex.ac.uk/linguistics/external/clmt/MTbook/HTML/book.html>. Accessed June 2012.
- Asia Online. 2012. Language Studio Newsletter.
<http://www.asiaonline.net/newsletters/201203.htm>

- Asia-online. Accessed June 2012. <http://www.asiaonline.net/newsletters/201203.htm>
- Atril. Accessed June 2012. Homepage of Déjà Vu X. www.atril.com.
- Austermühl, F. 2001. *Electronic Tools for Translators*. Manchester: St Jerome.
- Autodesk. 2012. Machine Translation at Autodesk.
http://translate.autodesk.com/productivity.html#prod_barall. Accessed March 2012.
- Beinborn, L. 2010. *Post-editing of statistical machine translation: A crosslinguistic analysis of the temporal, technical and cognitive effort*. Master of Science Thesis. Saarbrücken. Saarland University.
- Belam, J. 2003. "Buying up to falling down. A deductive approach to teaching post-editing". In *Proceedings of the 9th MT Summit. Workshop on Teaching Translation Technologies and Tools*. New Orleans. <http://www.mt-archive.info/MTS-2003-Belam.pdf> Accessed May 2012.
- Bennett, S. Gerber, L. 2003. "Inside commercial machine translation". In *Computers and Translation: A Translator's Guide*. Harold Somers, ed. Amsterdam and Philadelphia: Benjamins. 175-190.
- Beregovaya, O. Yanishevsky, A. 2010. "PROMT at PayPal: enterprise-scale MT deployment for financial industry content". In *Proceedings of the 9th Annual Conference of the AMTA*. Denver. Available from <http://amta2010.amtaweb.org/AMTA/papers/4-16-BeregovayaYanishevsky.pdf>. Accessed May 2012.
- Bier, K. Herranz, M. 2011. "MT experience at Sybase". Localization World Conference. Barcelona. Available from <http://www.slideshare.net/manuelherranz/loc-world2011-kbiermherranz-8730502> Accessed May 2012.
- Bowker, L. 2005. "Productivity vs Quality? A pilot study on the impact of translation memory systems". *Localisation Reader 2005-2006*: 133-140.
- Bowker, L. Ehgoetz, M. 2007. "Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation". Kenny, D. and Ryou, K., eds. 2007. *Across Boundaries: International Perspectives on Translation*. Newcastle-upon-Tyne: Cambridge Scholars Publishing. 209-224
- Bruckner, C. Plitt, M. 2001. "Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input". In *Proceedings of the 8th*

- EAMT Conference. MT Evaluation Workshop at Geneva*. Available from <http://www.mt-archive.info/MTS-2001-Bruckner.pdf>. Accessed June 2012.
- Brunette, L. Gagnon, C. Hine, J. 2005. "The Grevis Project. Revise or Court Calamity". *Across Languages and Cultures* 6 (1): 29-45.
- Carl, M. Dragsted, B. Elming, J. Hardt, D. Jakobsen, A. 2011. "The process of post-editing: a pilot study". In *Proceedings of the 8th international NLPSC workshop*. Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds). (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur: 131-142. Available from <http://www.mt-archive.info/NLPCS-2011-Carl-1.pdf>
- Carletta, J. 1996. "Assessing agreement on classification tasks: the kappa statistic". *Computational Linguistics*. Vol 22(2). Cambridge, MA: MIT Press. 249-254. Available from <http://acl.ldc.upenn.edu/J/J96/J96-2004.pdf>
- Christensen, T. Schjoldager, A. 2011. "The Impact of Translation-Memory (TM) Technology on Cognitive Processes: Student-Translators' Retrospective Comments in an Online Questionnaire". In *Proceedings of the 8th international NLPSC workshop*. Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds.) (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur. 119-130. Available from <http://www.mt-archive.info/NLPCS-2011-Christensen.pdf>
- Common Sense Advisory. Accessed June 2012.
<http://www.commonsenseadvisory.com/>
- Creswell, J. W. 2003. *Research Design. Qualitative, Quantitative and Mixed Methods Approaches*. Thousand Oaks. London. Delhi. Singapore: SAGE Publications.
- Creswell, J. W. Plano Clark, V. L. 2007. *Mixed Methods Research*. Thousand Oaks. London. Delhi. Singapore: SAGE Publications.
- CrossLang. Accessed June 2012. <http://www.crosslang.com/>
- De Almeida, G. O'Brien, S. 2010. "Analysing Post-Editing Performance: Correlations with Years of Translation Experience". In *Proceedings of the 14th Annual Conference of the EAMT*. St. Raphael. Available from <http://www.mt-archive.info/EAMT-2010-Almeida.pdf>. Accessed June 2012.
- De Palma, D. 2011. "Trends in Machine Translation". Common Sense Advisory. Available from www.commonsenseadvisory.com Accessed June 2012.

- De Palma, D. Kelly, N. 2009. "The Business case for Machine Translation". Common Sense Advisory. Available from www.common senseadvisory.com. Accessed June 2012.
- De Palma, D. Sargent, B. Bassetti, T. Beninato, R. 2008. "The price of Translation: a comprehensive analysis of pricing for globalization service buyers". Common Sense Advisory. Available from www.common senseadvisory.com. Accessed June 2012.
- De Sutter, N. 2012. "MT evaluation based on post-editing: a proposal". Depraetere, I. ed. *Perspectives on translation quality*. Berlin: Mouton de Gruyter.125-146.
- De Sutter, N. Depraetere, I. 2012. "Post-edited translation quality, edit distance and fluency scores: report on a case study". Presentation in *Journée d'études Traduction et qualité Méthodologies en matière d'assurance qualité*. Université Lille 3. Sciences humaines et sociales. Lille. Available from http://stl.recherche.univ-lille3.fr/colloques/20112012/DeSutter&Depraetere_2012_02_03.pdf Accessed May 2012.
- Depraetere, I. 2010. "What counts as useful advice in a university post-editing training context? Report on a case study". In *Proceedings of the 14th Annual EAMT Conference*. St. Raphael. Available from <http://www.mt-archive.info/EAMT-2010-Depraetere-2.pdf> Accessed June 2012.
- Dimitrova, B. 2005. *Expertise and Explicitation in the translation process*. Amsterdam and Philadelphia: Benjamins
- Doddington, G. 2002." Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics". In: *Proceedings of the 2nd International Conference on Human Language Technology*. San Diego. 138-145.
- Dove, C. Way, A. Johnson, D. 2011. "Quality above price and speed". TAUS Conference. Santa Clara. Video available at http://www.youtube.com/watch?v=eZDAD7Y_MHE Accessed June 2012.
- Dragsted, B. 2004. *Segmentation in Translation and Translation Memory Systems*. PhD Thesis. Copenhagen. Copenhagen Business School.
- EAMT. Accessed June 2012. Homepage of the European Association for Machine Translation www.eamt.org.
- Easton, V. McColl, J. 2012. Statistics Glossary. Accessed on-line <http://www.stats.gla.ac.uk/steps/glossary/index.html>

- Fiederer, R. O'Brien, S. 2009. "Quality and machine translation: a realistic objective?" *The Journal of Specialised Translation*. (11). Available from http://www.jostrans.org/issue11/art_fiederer_obrien.pdf Accessed June 2012.
- Fifer, M. T. 2007. *The Fuzzy Factor: An Empirical Investigation of Fuzzy Matching in the Context of Translation Memory Systems*. Master's thesis. University of Ottawa.
- Flournoy, R. Duran, C. 2009. "Machine Translation and Document Localization at Adobe: From Pilot to Production". In *Proceedings of the 12th MT Summit*. Ottawa. Available from <http://www.mt-archive.info/MTS-2009-Flournoy.pdf>. Accessed June 2012.
- Fundeu. 2012. Wikilengua. <http://www.wikilengua.org/index.php/Gerundio>
- García, I. 2005. "Long term memories: Trados and TM turn 20". *The Journal of Specialised Translation*. Issue 04. Available from http://www.jostrans.org/issue04/art_garcia.php. Accessed June 2012.
- García, I. 2006a. "Translators on translation memories: a blessing or a curse?" *Translation Technology and its Teaching*. Intercultural Studies Group. Universitat Rovira i Virgili. Tarragona. Available from http://isg.urv.es/library/papers/Garcia_Translators.pdf
- García, I. 2006b. "The Bottom Line: Does Text Reuse Translate into Gains in Productivity". *The International Journal of Technology, Knowledge and Society*. Vol. 2(1). Common Ground: 103-110.
- García, I. 2007. "Power shifts in web-based translation memory". *Machine Translation*, Vol. 21(1). Netherlands: Springer. 55-68.
- García, I. 2008. "Translating and Revising for Localisation: What do We Know?" *Perspectives Studies in Translatology*, 16(1-2). London: Routledge: 49-60.
- García, I. 2009. "Beyond Translation Memory: Computers and the Professional Translator". *Journal of Specialised Translation*, Issue 12. Available from http://www.jostrans.org/issue12/art_garcia.pdf. Accessed June 2012.
- García, I. 2010. "Is Machine Translation Ready Yet?" *Target*. Vol. (22-1). Amsterdam and Philadelphia: Benjamins. 7-21
- García, I. 2011. "Translating by post-editing: Is it the way forward?" *Machine Translation*, Vol. 25(3). Netherlands: Springer. 217-237
- Gow, F. 2003. "Extracting useful information from TM databases." *Localisation Reader* 2004-2005: 41-44.

- Greenacre, M. 2008. *La práctica del análisis de correspondencias*. Madrid: Fundación BBVA, Madrid (Spanish translation of Correspondence Analysis in Practice, Second Edition). Available from <http://www.fbbva.es/TLFU/tlfu/esp/publicaciones/libros/fichalibro/index.jsp?codigo=300>
- Groves, D. Schmidtke, D. 2009. "Identification and analysis of post-editing patterns for MT". In *Proceedings of the 12th MT Summit*. Ottawa. Available from <http://www.mt-archive.info/MTS-2009-Groves.pdf>. Accessed June 2012.
- Guerberof, A. 2008. *Productivity and Quality in Machine Translation and Translation Memory outputs*. Masters Dissertation. Tarragona. Universitat Rovira i Virgili. Also available from https://docs.google.com/open?id=0B_-cVNsnGLfvMjc5MDc3ZWYtZjRiZC00NWZLTlhMTetNTEzOWIwODY0MDc1
- Guerra Martínez, L. 2003. *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output*. Masters Dissertation. Dublin. Dublin City University.
- He, Y. Ma, Y. Roturier, J. Way, A. van Genabith, J. 2010b. "Improving the Post-Editing Experience using Translation Recommendation: A User Study". In *Proceedings of the 9th Annual AMTA Conference*. Denver: 247-256. Available from <http://doras.dcu.ie/15803/>
- He, Y. Ma, Y. Van Genabith, J. Way, A. 2010a. "Bridging SMT and TM with Translation Recommendation". In *Proceedings of the 48th Annual Meeting of ACL*. Uppsala: 622-630. Available from <http://www.mt-archive.info/ACL-2010-He.pdf> Accessed June 2012.
- HiSoft. Accessed June 2012. <http://www.hisoft.com/>
- Hogan, C. Allen, J. 2000. "Toward the development of a post-editing module for raw machine translation output: A controlled language perspective". In *Proceedings of the Third International Controlled Language Applications Workshop*. Seattle, Washington. Available from <http://www.oocities.org/mtpostediting/> Accessed June 2012.
- Holyk, G. 2008. "Questionnaire Design". *Encyclopedia of Survey Research Methods*. Thousand Oaks. London. Delhi. Singapore: SAGE Publications.

- Hutchins, W. J. 1995. "Machine Translation: A brief history". *Concise history of the language sciences: from the Sumerians to the cognitivists*. E. F. K. Koerner and R. E. Asherd, eds. Oxford. Pergamon Press: 431-445.
- Hutchins, W. J. 2001. "Machine Translation over fifty years". *Histoire, Epistemiologie, Langage*. Tome XXII, fasc. 1: 7-31. Available from <http://www.hutchinsweb.me.uk/HEL-2001.pdf>
- Koehn, P. 2012a. *Moses. A statistical Translation System. User Manual and Code Guide*. Available from <http://www.statmt.org/moses/manual/manual.pdf>
- Koehn, P. 2012b. "What is a Better Translation?" Reflections on Six Years of Running Evaluation Campaigns". *Tralogy* [En ligne session 5- Quality in Translation / La qualité en traduction, mis à jour le 31/01/2012] Available from <http://homepages.inf.ed.ac.uk/pkoehn/publications/tralogy11.pdf>
- Koehn, P. Hoang, H. Birch, A. Callison-Burch, C. Federico, M. Bertoldi, N. Cowan, B. Shen, W. Moran, C. Zens, R. Dyer, C. J. Bojar, O. Constantin, A. and Herbst, E. 2007. "Moses: Open source toolkit for statistical machine translation". In *Proceedings of the 45th Annual Meeting of the ACL Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics. Prague: 177-180.
- Krings, H. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*. G. S. Koby, ed. Ohio. Kent State University Press.
- Künzli, A. 2006. "Translation revision - A study of the performance of ten professional translators revising a technical text". *Insights into specialized translation*. Maurizio Gotti and Susan Sarcevic (eds). Bern & Frankfurt: Peter Lang: 195-214.
- Künzli, A. 2007. "Translation Revision. A study of the performance of ten professional translators revising a legal text". *Doubts and Directions in Translations Studies. Selected contributions from EST Congress Lisbon 2004*. Gambier, Y. Shlesinger, M. Stotlze, R. (eds.). Amsterdam and Philadelphia: Benjamins. 115-126.
- Lagoudaki, E. 2006. "Translation Memories Survey 2006: Users' perceptions around TM use". *Translating and the Computer*, 28. London: Aslib. Also available from <http://mt-archive.info/Aslib-2006-Lagoudaki.pdf>. Accessed June 2012.
- Lagoudaki, E. 2008. "The value of Machine Translation for the Professional Translator". In *8th AMTA Conference*. Hawaii: 262-269. Available from http://www.amtaweb.org/papers/3.04_Lagoudaki.pdf

- Language Weaver. Accessed June 2012. <http://www.sdl.com/en/language-technology/landing-pages/languageweaver/>
- Littell, R. C. Stroup, W. W. Freund, R. J. 2005. *SAS® for Linear Models*. Fourth Edition. Cary, NC: SAS Institute Inc.
- Loffler-Laurian, A. M. 1983. “Pour une typologie des erreurs dans la traduction automatique”. [Towards a typology of errors in Machine Translation], *Multilingua* (2): 65–78.
- Loffler-Laurian, A. M.: 1986. “Post-édition rapide et post-édition conventionnelle: deux modalités d’une activité spécifique”. [Rapid post-editing and conventional post-editing: two modalities of a specific activity]. *Multilingua* (5): 225-229.
- Lorenzo, M. P. 2002. “Competencia revisora y traducción inversa.” [Translation Competence and Reverse Translation]. *Cadernos de Tradução* (10): 133-166. Also available from <http://www.periodicos.ufsc.br/index.php/traducao/article/view/6148/5706> Accessed June 2012.
- Martín-Mor, A. 2011. *La interferència lingüística en entorns de Traducció Assistida per Ordinador*. [Linguistic interference in Computer Assisted Tools] Doctoral Thesis. Barcelona. Universitat Autònoma de Barcelona.
- Melamed, I. D. Green, R. Turian, J. P. 2003. “Precision and Recall of Machine Translation”. In *Proceedings of HLT-NAACL 2003*. Edmonton: 61-63. Also available from <http://nlp.cs.nyu.edu/pubs/papers/hlt03eval.pdf>
- Mesa, B. 2011. “Explicitation in translation memory-mediated environments. Methodological conclusions from a pilot study”. *Translation & Interpreting*. Vol (3-1): 44-57. Available from <http://transint.org/index.php/transint/article/viewFile/137/84>. Accessed June 2012.
- Moorkens, J. 2011. “Translation Memories guarantee of consistency: Truth or fiction?” *Translation and the Computer 17 & 18*. London: Aslib. Available from <http://www.cngl.ie/drupal/sites/default/files/papers3/Aslib-2011-Moorkens.pdf>.
- Morse, J. M. 2003. “Principles of Mixed Methods and Multimethod Research Design”. In Tashakkori, A and Teddlie, C. (eds.). *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks. London. Delhi. Singapore: SAGE Publications. 189-208.
- Morse, J. M. Niehaus, L. 2009. *Mixed Method Design. Principles and Procedures*. California: Left Coast Press.

- Mossop, B. 2001. 2007a. *Editing and Revising for Translators*. Manchester: St. Jerome Publishing.
- Mossop, B. 2007b. “Empirical studies of revision: what we know and need to know”. *Journal of Specialised Translation*. Issue 8. Available from http://www.jostrans.org/issue08/art_mossop.pdf. Accessed June 2012.
- O’Brien, S. 2002. “Teaching post-editing: A proposal for course content”. In *Proceedings for the 6th Annual EAMT Conference. Workshop Teaching machine translation*. Manchester: 99-106. Available from <http://mt-archive.info/EAMT-2002-OBrien.pdf>. Accessed June 2012.
- O’Brien, S. 2006a. “Methodologies for Measuring Correlations between Post-Editing Effort and Machine Translatability” *Machine Translation*. Netherlands: Springer. 37-58.
- O’Brien, S. 2006b. “Eye-tracking and Translation Memory Matches” *Perspectives: Studies in Translatology*. 14 (3): 185-205
- O’Brien, S. 2011. “Towards predicting post-editing productivity”. *Machine Translation*, Vol. 25(3). Netherlands: Springer. 197-215
- O’Brien, S. 2012. “Towards a Dynamic Quality Evaluation Model for Translation”. *Journal of Specialised Translation*. (17) Available from http://www.jostrans.org/issue17/art_obrien.pdf
- O’Leary, Z. 2004. *The Essential Guide to Doing Research*. SAGE Publications. Thousand Oaks. London. Delhi. Singapore: SAGE Publications. 189-208.
- O’Leary, Z. 2010. *The essential Guide to doing your research project*. Thousand Oaks. London. Delhi. Singapore: SAGE Publications.
- O’Brien, S. O’Hagan, M. Flanagan, M. 2010. “Keeping an eye on the UI design of Translation Memory: How do translators use the 'concordance' feature?” In: *European Conference on Cognitive Ergonomics*. Delft: 25-28. Available from <http://doras.dcu.ie/16693/>
- Offersgaard, L. Povlsen, C. Almsten, L. Maegaard, B. 2008. “Domain specific MT use”. In *Proceedings of 12th EAMT Conference*. Hamburg: 150-159. Available from <http://www.mt-archive.info/EAMT-2008-Offersgaard.pdf>
- Olivier, J. Hand, D. 1996. “Introduction to minimum decoding inference”. Tech Report 205. Department of Computer Science. Monash University, Clayton, Vic. 3168. Australia.

- Paladini, P. 2011. "Translator's productivity increase at CA Technologies". Localization World Conference. Barcelona. PPT available from www.localizationworld.com/lwbar2011/presentations/files/D1.pptx Accessed June 2012.
- Papineni, K. Roukos, S. Ward, T. Zhu, W.J. 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: 311-318. Also available from <http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf>. Accessed June 2012.
- 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. Prague: 7-16. Available from <http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf> Accessed June 2012.
- Quah, C. K. 2006. *Translation and Technology*. New York: Palgrave MacMillan.
- Quirk, C. Menenzes, A. Cherry, C. 2005. "Dependency Treelet Translation: Syntactically Informed Phrasal SMT". In *Proceedings of the 43rd Annual Meeting of ACL*. Ann Arbor: 271-279. Available from <http://www.aclweb.org/anthology-new/P/P05/P05-1034.pdf>. Accessed June 2012.
- Real Academia Española. Asociación de Academias Americanas. 2009. *Nueva Gramática de la Lengua Española*. Ignacio Bosque, ed. Madrid: Espasa Libros, S.L.U.
- Ribas, C. 2007. *Translation Memories as vehicles for error propagation. A pilot study*. Minor Dissertation. Tarragona. Universitat Rovira i Virgili.
- Rieche, A. 2004. *Memória de tradução: auxílio ou empecilho?* [Translation Memory: Aid or handicap?]. Masters Dissertation. Rio de Janeiro. Pontíficia Universidade Católica do Rio de Janeiro. Available in Portuguese from http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=4974@2
- Roukos, S. Rojas, F. Ming Xu, J. Pont Nesta, S. Martínez Corriá, A. Chapman, H. Vohra, S. 2011. "The value of post-editing: IBM case study". Localization World Conference. Barcelona. Available from www.localizationworld.com/lwbar2011/presentations/files/E6.ppt
- Schäffer, F. 2003. "MT post-editing: How to shed light on the 'unknown task' Experiences made at SAP". In *Proceedings for the 8th International Workshop of*

the EAMT and the 4th Controlled Language Applications Workshop. Dublin.

Available from <http://www.mt-archive.info/CLT-2003-Schaefer.pdf>. Accessed June 2012.

Schäler, R. 1998. "The Problem with Machine Translation". *Unity in Diversity: Recent Trends in Translation Studies*. Bowker et al., eds. Manchester: St. Jerome. 151-156.

SDL Automated Translation. Accessed June 2012.

<http://www.translationzone.com/en/translator-solutions/automated-translation/>

SDL. Homepage of SDL Trados 2007. Accessed June 2012. <http://www.trados.com/en/>

Senez, D. 1998. "The Machine Translation Help Desk and the Post-Editing Service".

Terminologie et Traduction 1: 289-295. Available from <http://www.mt-archive.info/T&T-1998-Senez.pdf>. Accessed June 2012

Siegel, S. Castellan, N. J. 1988. *Nonparametric statistics for behavioral sciences*. New York: McGraw-Hill.

Snover, M. Dorr, B. Schwartz, R. Micciulla, L. Makhoul, J. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation". In *Proceedings of 7th Annual AMTA Conference*. Washington DC: 223-231. Available from <http://mt-archive.info/AMTA-2006-Snover.pdf>. Accessed June 2012.

Somers, H. 2003. "Translation Memory Systems". In *Computers and translation: A translator's guide*. Harold Somers, ed. Amsterdam and Philadelphia: Benjamins. 31-47

Specia, L. 2011. "Exploiting Objective Annotations for Measuring Translation Post-editing Effort". In *Proceedings of the 15th Annual EAMT Conference*. Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste, eds. Leuven. 73-80. Available from <http://www.mt-archive.info/EAMT-2011-Specia.pdf> Accessed June 2012.

Specia, L. Cancedda, N. Dymetman, M. Turchi, M. Cristianini, N. 2009a. "Estimating the sentence-level quality of machine translation systems". In *Proceedings of the 13th Annual Conference of the EAMT*. Barcelona: 28-35. Available from http://clg.wlv.ac.uk/papers/Specia_EAMT2009.pdf Accessed June 2012.

Specia, L. Saunders, C. Turchi, M. Wang, Z. Shawe-Taylor, J. 2009b. "Improving the confidence of machine translation quality estimates". In *Proceedings of the 12th MT Summit*. Ottawa: 136 – 143. Available from <http://eprints.pascal-network.org/archive/00005490/01/MTS-2009-Specia.pdf> Accessed June 2012.

- Star Group. Homepage of Star Transit. Accessed June 2012. www.star-group.net/star-www/description/transit/star-group/eng/star.html.
- Statsoft. Accessed June 2012. www.statsoft.com
- Systran. Homepage of the language translation software provider. Accessed June 2012. www.systran.co.uk/.
- Tashakkori, A. Teddlie, C. 2009. *Foundations of Mixed Methods Research*. Thousand Oaks. London. Delhi. Singapore: SAGE Publications
- Tatsumi, M. 2009. "Correlation between automatic evaluation metric scores, post-editing speed, and some other factors". In *Proceedings of the 12th MT Summit*. Ottawa: 332-339. Available from <http://www.mt-archive.info/MTS-2009-Tatsumi.pdf> Accessed June 2012.
- Tatsumi, M. 2010. *Post-Editing Machine Translated Text in A Commercial Setting: Observation and Statistical Analysis*. PhD Thesis. Dublin City University. Available from <http://doras.dcu.ie/16062/> Accessed March 2012.
- Tatsumi, M. Roturier, J. 2010. "Source text characteristics and technical and temporal post-editing effort: What is their relationship?" in *Proceedings of the 2nd Joint EM+/CNGL workshop "Bringing MT to the user: research on integrating MT in the translation industry"*. Ventsislav Zhechev, ed. Denver: 43-51. Available from <http://www.mt-archive.info/JEC-2010-Tatsumi.pdf> Accessed June 2012.
- TAUS. 2010a. "TAUS Research. Results post-editing survey". Available from <http://www.translationautomation.com/articles/>
- TAUS. 2010b. "Postediting in Practice. A TAUS Report". Available from <http://www.translationautomation.com/articles/>
- TAUS. 2012. "Moses: Commodity Creates Opportunity". Available from <http://www.translationautomation.com/technology/moses-commodity-creates-opportunity.html>. Accessed May 2012.
- TAUS. Accessed June 2012. www.translationautomation.com/.
- Teixeira, C. 2011. "Knowledge of Provenance and its Effects on Translation Performance in an Integrated TM/MT Environment". In *Proceedings of the 8th international NLPSC workshop*. Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen, eds. (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur: 107-118. Available from <http://www.mt-archive.info/NLPCS-2011-Teixeira.pdf>

- Torres-Hostench, O. Biau Gil, R. Cid Leal, P. Martín Mor, A. Mesa-Lao, B. Orozco, M. Sánchez Gijón, P. 2010. "TRACE: measuring the impact of CAT tools on translated texts". *Linguistic and Translation Studies in Scientific Communication*. Gea et al., eds. Peter Lang: 255-276
- Turian, J. Shen, L. Melamed, I. D. 2003. "Evaluation of Machine Translation and Its Evaluation". In *Proceedings of the 9th the MT Summit*. New Orleans: 386-393. Available from <http://nlp.cs.nyu.edu/pubs/papers/turian-summit03eval.pdf>. Accessed June 2012.
- Vasconcellos, M. 1986. "Post-Editing on Screen: Machine Translation from Spanish into English". In *Translating and the Computer 8: A Profession on the Move*. Catriona Picken, ed. London: Aslib. 133-146.
- Vasconcellos, M. 1992. "What do we want from MT?" *Machine Translation*. Vol 7(4). Netherlands: Springer. 293-301.
- Vasconcellos, M. 1993. "Machine Translation. Translating the languages of the world on a desktop computer comes of age". *BYTE*. McGraw Hill: 153-164
- Vasconcellos, M. 1989. "Cohesion and coherence in the presentation of machine translation products". *Georgetown University Round Table on Languages and Linguistics*. James E. Alatis (ed). Washington, D.C.: Georgetown University Press. 90-105. Available from <http://www.mt-archive.info/GURT-1989-Vasconcellos.pdf>
- Vasconcellos, M. León, M. 1985. "SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization". *Computational Linguistics*. Vol. 11 (2-3): 122-136. Available from <http://acl.ldc.upenn.edu/J/J85/J85-2003.pdf>.
- Vashee, K. 2012a. "Need for automated quality measurement" Entry from blog <http://kv-emptypages.blogspot.com/2010/03/need-for-automated-quality-measurement.html>
- Vashee, Kirti. <http://www.scoop.it/t/automated-translation-mt-trends?page=2>
- Vilanova, S. 2006. *L'impacte de les memòries de traducció sobre el text d'arribada: interferències i trets lingüístics*. [The impact of translation memories on the target text: interferences and linguistic traits] Masters Dissertation. Tarragona. Universitat Rovira i Virgili.
- Wagner, E. 1985. "Rapid post-editing of Systrans". In *Translating and the Computer 5: Tools for the trade*. Veronica Lawson, ed. London: Aslib. 199-213

- Wagner, E. 1987. "Post-editing: Practical considerations". In *ITI Conference I: The Business of Translating and Interpreting*. Catriona Picken, ed. London: Aslib. 71-78.
- Wallis, J. 2006. *Interactive Translation vs. Pre-translation in the Context of Translation Memory Systems: Investigating the effects of translation method on productivity, quality and translator satisfaction*. Thesis for Master in Translation Studies. Ottawa. University of Ottawa. Available from <http://www.localisation.ie/resources/Awards/Theses/Thesis%20-%20Julian%20Wallis.pdf>. Accessed June 2012.
- Yamada, M. 2011a. "The effect of translation memories on productivity". *Translation Research Projects 3*. Intercultural Studies Group. Tarragona. Universitat Rovira i Virgili. 63-73
- Yamada, M. 2011b. *Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process*. PhD Thesis. Tokyo. Rikkyo University.

Appendix A

Glossary

Estimation and estimate

Estimation is the process by which sample data are used to indicate the value of an unknown quantity in a population. An estimate is the particular value of an estimator that is obtained from a particular sample of data and used to indicate the value of a parameter (Easton and McColl 2012).

Explanatory and response variables

Explanatory variable is a synonym for independent variable. This is the variable that tries to explain or predict changes in the values of another variable (response or dependent variable).

Confidence intervals

The confidence intervals for the mean give a range of values around the mean where the "true" (population) mean is expected to be located (with a given level of certainty). For example, if the mean is 23, and the lower and upper limits of 95 percent confidence interval are 19 and 27 respectively, then the probability that the population mean is greater than 19 and lower than 27 is 95 percent, i.e. $m 0.95$. If the confidence is greater, then the interval would become wider thereby increasing the "certainty" of the estimate, and vice versa. The width of the confidence interval depends on the sample size and on the variation of data values. The larger the sample size, the more reliable the mean. The larger the variation, the less reliable the mean is. (StatSoft 2012)

Continuous and discrete variables

A continuous variable can assume an infinite number of values within an interval, for example *Height*. A discrete variable describes a finite or countable set of values, for example, *Male* or *Female*.

Kappa coefficient

In statistics, the Kappa coefficient is a measurement to indicate inter-rater or inter-annotator agreement:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Where $P(A)$ is the proportion of times that the raters agree and $P(E)$ is the proportion of times that they would be expected to agree by chance. When there is no agreement other than that which would be expected by chance, K is zero. When there is total agreement, K is one. (Carletta 1996).

Kruskal-Wallis analysis of variance

The Kruskal-Wallis is a non-parametric test used to compare three or more independent samples belong to the same population based on the medians. The null hypothesis is that the medians are the same as that for the population, with the statistical text checking to see if they are close enough, allowing for natural variation in samples. The data are assumed to be at least ordinal and there is a single underlying continuous distribution (Siegel and Castellan 1988).

Fuzzy match

In a Translation memory system, a Fuzzy match is a partial proposal that the translation system gives if no full match is available (a full match is an exact correspondence between the old and new source called 100 percent match). Different levels of fuzzy matches are calculated with an algorithm that uses character strings similarities and establishes partial correspondences between the source sentences based on syntactic structures. Different tools use different algorithms. The fuzzy match value is normally expressed as a percentage. This percentage represents the characters that were translated with a less than perfect translation memory match (100 percent). Therefore, a 95 percent, match, for example, is considered a high percentage match, that is, the translation proposed is deemed to be very close to the new source text.

LISA and the LISA QA Model

The Localization Industry Standards Association (LISA) was an association dedicated to the creation and implementation of standards for the localization industry. Through their OSCAR interest group they were responsible for the creation of various standards such as TermBase eXchange (TBX), TBX-Basic, Translation Memory Exchange (TMX), Segmentation Rule eXchange (SRX), Global Information management Metrics eXchange (GMX), XML text memory (xml:tm) and Term Link. They were also responsible for the creation of the LISA QA Model for the evaluation of localized project quality. The quality metrics and procedures in the QA Model were the results of collaboration between LISA members, localization services providers, software and hardware developers, and end-users. Although LISA closed in 2011, the review form or QA model originally created by them is still widely used in the localization industry.

LISA defines different type of errors. These are Mistranslation, Accuracy, Terminology, Language, Style, Country, Consistency and Format. Mistranslation refers to the incorrect understanding of the source text; Accuracy to omissions, additions, cross-references, headers and footers and not reflecting the source text properly; Terminology to glossary adherence; Language to grammar, semantics, spelling, punctuation; Style to adherence to style guides; Country to country standards and local suitability; Consistency to coherence in terminology across the project and Format to correct use of tags, correct character styles, correct footnotes translation, hotkeys not duplicated, correct flagging, correct resizing, correct use of parser, template or project settings file.

The errors found are then assigned a severity level that can be Minor, Major or Critical. All errors are weighted according to these categories. For example, an error classified as Minor weighs one point, if classified as Major, five points, and finally if it is deemed to be Critical it is penalized with the total amount of allowed errors plus one.

Fuzzy match and MT post-editing pricing

Within the localization industry, there is a relative agreement about pricing of new words and translation memory segments. Normally, translation is paid per word (in some cases per line or per hour is used) and fuzzy matches are paid according to a percentage of that word rate. For example, if the TM determines that one segment constitutes a 100 percent match, then this segment tends to be paid between 20 and 30

percent of the total word price, therefore assuming that there is a 70 percent saving and this should correspond to a high reduction in the translation effort. Table 86 shows the level of fuzzy matches, the percentage paid, and the assumed savings. The figures might vary depending on the companies, these are orientative values.

Level of match	% of word paid	% of saving
100% match & repetitions	20-30%	80-70%
95-99% fuzzy match	20-30%	80-70%
75% -94%	60%	40%
0-74%	100%	0%

Table 86: Pricing for fuzzy matches a percentage of full word rate

Common Sense Advisory has published a report on the pricing of translation (De Palma et al. 2008) where they establish that most agencies pay 35 percent of the full word rate for repetitions and 100 percent matches, and between 50-65 percent for fuzzy matches in the 80-90 percent category (2008: 11). The average figures are in line with the figures in the table above.

In recent years, there has been more information available about how post-editing is paid in the localization industry. This is a much debated topic in localization conferences and publications. Due to the nature of this task and the different variables, there is no general agreement on how the MT segments should be paid. TAUS has published results from a survey carried out in 2010 (TAUS 2010a) that indicates that 52.2 percent of Language Service Providers offer post-editing services, and that they pay post-editors either a price per word (40 percent of companies) as a fuzzy match or per hour (34 percent of companies). In a report, also from TAUS, (2010b) we find that there are a variety of pricing models, the most popular ones being:

1. Paying for post-editing as fuzzy segment matches
2. Paying a fee based on time spent.

The variations on the per word/segment rate include:

- Lower than fuzzy match rate: i.e. between 15 and 25 percent
- A per-word discount on the price
- A percentage of the no-match word rate
- 50 percent of usual human translation rate
- A rate based on productivity

Common Sense Advisory has also produced a report (De Palma and Kelly 2009) showing that LSPs pay from 45 to 90 percent of the cost of human translators for post-edited MT output.

Overcorrection

According to the Random House Dictionary, overcorrection is “the correction beyond what is needed or customary, especially when leading to error; over adjustment”. We have used it in this study for those edits performed by translators that did not necessarily address an error.

Poisson distribution

The Poisson distribution is also sometimes referred to as the distribution of number of events. An example of Poisson distributed variables is number of accidents per person. It is defined as:

$$f(k; x) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where

λ is a parameter that represents the expected frequency of the event

Post-editing

To post-edit is defined as “to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)” (Allen 2003: 296) or “revising the output of a machine translation program”, where “revising” means “the process of checking a *draft translation* for *errors* and making appropriate *amendments*” (Mossop 2001: 168-169, italics in original). Here we assume that in post-editing the “draft translation” is the “MT output”.

Preferential changes

In this study, a preferential change is defined as those edits that do not involve the fixing of errors or changes that are not fully justifiable within the context of the instructions given to the participants. This definition is included in the instructions given to translators and reviewers.

Processing speed

Processing speed is expressed as the number of source words processed per minute in each of the three categories (No match, Fuzzy match and MT match). The processing speed is thus measured in words per minute. Krings (2001) and O'Brien (2006b) use processing speed in order to measure productivities and different correlations in text types.

Processing time

Processing time is the total time used to process segments from each of the three categories: No match, MT match and Fuzzy match. The processing time is measured in minutes. Krings uses processing time when discussing temporal post-editing and he defines it as the “time used by subjects working on a specific task” (2001:276). He operationalizes it as the “time between taking up and laying aside the text, minus possible non-task-related interruptions” (2001:276) because part of his experiment was done on paper. In our case, we will automatically record time invested by the post-editor by means of an on-line post-editing tool. When the post-editor presses the Next button the time starts being recorded until he or she presses the Next button again to go to the following segment.

Productivity gain

The productivity gain is the relationship existing between the processing speed of one post-editor translating a new segment and the processing speed of that same post-editor when using the aid of a tool, TM or MT, for the same amount of words. This gain is expressed as a percentage value. This concept is similar to the Relative post-editing effort define by Krings (2001) but in our case we use a percentage value.

Reviewer

We have used the term “reviewer” exclusively to refer to the participants that evaluated the post-edited text that the translators had worked on. These reviewers are also translators and they perform both types of tasks as part of their daily activities.

Translator and post-editor

We have used the term “translator” and “post-editor” to refer to the participants that carried out the assignment as they had to translate new text and to post-edit translation memory and machine-translated output. We have also used both terms interchangeably to refer to the same participants as post-editing tasks are normally performed by professional translators (TAUS 2010a).

Translation Edit Rate (TER)

TER (Snover et al. 2006) is an automatic score that reflects the number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references. In this project, we have used TER to compare the edits made to the Fuzzy and MT matches by the translators. See section Translation Edit Rate (TER) for more information on the methodology followed.

Variable of interest

A variable of interest is the variable that constitutes the focus of the study. For example, if trying to find the average height of the population, the variable of interest is *Height*.

Appendix B

This appendix contains the communication with the translators and can be found here:

<http://kcy.me/9t6u>

Appendix C

This appendix contains the instructions sent to the translators and the reviewers and can be found here:

<http://kcy.me/9t6w>

Appendix D

This appendix contains the questionnaires sent to the translators here:

<http://kcy.me/9t6z>

And the one sent to the reviewers here:

<http://kcy.me/9t6y>

Appendix E

This appendix contains the LISA QA form slightly modified for the assignment and can be found here:

<http://kcy.me/9t74>

Appendix F

This appendix contains the descriptive processing speed values per translator and can be found here:

<http://kcy.me/9t75>

Appendix G

This appendix contains the number of errors per translator and can be found here:

<http://kcy.me/9t76>

Appendix H

This appendix contains the error indicator per translator and can be found here:

<http://kcy.me/9t77>

Appendix I

This appendix contains the translators' transcripts of debriefings as well as a summary of values per translator. It can be found here:

<http://kcy.me/9t78>

Appendix J

This appendix contains the reviewers' transcripts of debriefings and can be found here:

<http://kcy.me/9t79>