THE IMPACT OF METADATA ON TRANSLATOR PERFORMANCE: HOW TRANSLATORS WORK WITH TRANSLATION MEMORIES AND MACHINE TRANSLATION.

Carlos da Silva Cardoso Teixeira

Dipòsit Legal: T 264-2015

Carlos da Silva Cardoso Teixeira

THE IMPACT OF METADATA ON TRANSLATOR PERFORMANCE: HOW TRANSLATORS WORK WITH TRANSLATION MEMORIES AND MACHINE TRANSLATION

DOCTORAL THESIS

Submitted in partial fulfilment of the requirements for a Double Doctorate in
Translation and Intercultural Studies (URV)
and in Translation Studies (KU Leuven)

Co-supervised by:

Dr Anthony Pym
Facultat de Lletres
Dept. of English and German Studies
Universitat Rovira i Virgili
Tarragona, Spain

Dr Reine Meylaerts
Faculteit Letteren
Translation Studies Research Unit
Katholieke Universiteit Leuven
Leuven, Belgium

UNIVERSITAT ROVIRA I VIRGILI

KU LEUVEN

2014

UNIVERSITAT
ROVIRA I VIRGILI

Professor Anthony Pym
Intercultural Studies Group
URV. Avda. Catalunya 35
43002 Tarragona, Spain
http://isg.urv.es/

October 18, 2014

I hereby certify that the present study *The impact of metadata on translator performance: How translators work with translation memories and machine translation*, presented by Carlos da Silva Cardoso Teixeira for the award of the degree of Doctor, has been carried out under the supervision of myself at the Universitat Rovira i Virgili with co-supervision by Professor Reine Meylaerts at KU Leuven.

The research and the thesis fulfill all the conditions for the award of an INTERNATIONAL DOCTORATE, in accordance with current Spanish legislation.

UNIVERSITAT
ROVIRA I VIRGILI
DEPARTAMENT D'ESTUDIS ANGLESOS I ALEMANYS

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

President
European Society for
Translation Studies

FACULTEIT LETTEREN / FACULTY OF ARTS
ONDERZOEKSEENHEID VERTAALWETENSCHAP
RESEARCH UNIT TRANSLATION STUDIES
BLIJDE INKOMSTSTRAAT 21 PB3310
B-3000 LEUVEN

KU LEUVEN

Leuven, October 20, 2014

I hereby certify that the present study The impact of metadata on translator performance: How translators work with translation memories and machine translation, presented by Carlos da Silva Cardoso Teixeira for the award of the degree of Doctor, has been carried out under the co-supervision of myself at KU Leuven and Professor Anthony Pym at the Universitat Rovira i Virgili.

1425
KU Leuven
Faculteit Letteren – Faculty of Arts
ONDERZOEKSEENHEID
VERTAALWETENSCHAP
RESEARCH UNIT
TRANSLATION STUDIES
Blijde Inkomststraat 21 bus 3310
3000 LEUVEN, BELGIE

Reine Meylaerts
Chair Research Unit Translation Studies
KU Leuven, Faculty of Arts
Blijde Inkomststraat 21 - PB3310
3000 Leuven, Belgium
Reine.meylaerts@kuleuven.be
tel. +32 16 32 48 48

Abstract

This thesis investigates whether and how translation metadata affect translator performance in a workflow that combines suggestions from translation memories and machine translation. The study is based on a translation process experiment with 10 professional translators working from English into Spanish in a workplace setting.

The keystroke logging tools Inputlog and MTeval allowed for the collection of data on translation times and typing effort. BB FlashBack was used for screen and face recording. A Tobii eye tracker was used to identify how the translators shifted their attention between different parts of the screen. The final translations were assessed for quality by two professional reviewers using an error-score system. Finally, interviews were used for eliciting opinions from participants about certain aspects of their performance.

The quantitative data were analysed with mixed-effects linear(ised) regression models. The results show that translation metadata affect translation time and typing effort, and that the effects vary according to the type of translation suggestion (exact matches, fuzzy matches, machine translation). As a complementary finding, the current study identified no significant correlation between the translators' performances while typing and their performances while translating.

The qualitative data obtained from the interviews show a mismatch between the translators' perceived performance and their measured performance. They tended to prefer an environment with translation suggestions and metadata, even when this environment did not correspond with better performance. The translators mentioned metadata as a helpful feature in the translation tool, among other reasons because metadata help them adapt their translation strategies more easily according to the suggestion type. Task familiarity was also identified as an important factor affecting translators' perceptions.

The results obtained in this study suggest the need to advance research on how translators interact with translation tools, with a view to increase not only productivity but also job satisfaction. This thesis is expected to have also contributed to the field in terms of the methodology of workplace studies, by presenting some challenges and solutions. An important lesson is the need to find an optimal balance between ecological validity and data validity when conducting translation experiments in realistic scenarios.

Resum

Aquesta tesi investiga si les metadades de traducció afecten el rendiment del traductor en un flux de treball que combina propostes de traducció automàtica i de memòries de traducció. L'estudi es basa en un experiment sobre el procés de traducció de 10 traductors professionals que treballen de l'anglès a l'espanyol en un entorn laboral real.

Les eines de captura de teclat Inputlog i MTeval proporcionen les dades de temps de traducció i esforç de tecleig. BB FlashBack permet gravar les activitats en pantalla i les cares dels traductors. El sistema de seguiment ocular Tobii ajuda a identificar com els traductors distribueixen l'atenció entre les diferents parts de la pantalla. Dos revisors professionals avaluen la qualitat de les traduccions fent servir un sistema de puntuació d'errors. Finalment, mitjançant entrevistes, es recull l'opinió dels participants sobre certs aspectes de la seva actuació.

Les dades quantitatives s'analitzen amb models de regressió lineal (o linealitzada) d'efectes mixtos. Els resultats mostren que les metadades de traducció afecten el temps de traducció i l'esforç de tecleig a diferents nivells segons el tipus de proposta de traducció (coincidències exactes, coincidències parcials, traducció automàtica). Com a resultat complementari, aquest estudi no ha identificat una correlació significativa entre el rendiment dels traductors al teclejar i el seu rendiment al traduir.

Les dades qualitatives obtingudes a partir de les entrevistes mostren una manca de correspondència entre el rendiment mesurat i el rendiment percebut pels traductors. Els traductors solen preferir un entorn amb propostes de traducció i metadades, fins i tot quan aquest entorn no es correspon amb un millor rendiment. Els traductors consideren les metadades una característica útil en l'eina de traducció, entre altres raons, perquè els ajuden a adaptar més fàcilment les seves estratègies de traducció segons el tipus de proposta. La familiaritat amb la tasca també s'identifica com un factor important que afecta les percepcions dels traductors.

Els resultats obtinguts en aquest estudi suggereixen una necessitat d'avançar en la investigació sobre la interacció entre els traductors i les eines de traducció, amb la finalitat d'augmentar no només la productivitat sinó també la satisfacció laboral. Aquesta tesi espera contribuir també a la metodologia de la recerca en entorns laborals. Els reptes i solucions que presenta reafirmen la necessitat de trobar un equilibri entre la validesa ecològica i la validesa de les dades quan es realitzen experiments en escenaris realistes.

Samenvatting

Dit proefschrift onderzoekt of en hoe metadata over vertalen de prestatie van de vertaler beïnvloeden in een workflow die suggesties van vertaalgeheugens en machinevertaling combineert. Het proefschrift is een experimentele studie van het vertaalproces van tien professionele vertalers die vertalen van het Engels naar het Spaans in de werkplek.

De keystroke logging software Inputlog en MTeval werden gebruikt om data over vertaaltijd en typinspanning te verzamelen, en er werden scherm- en gezichstopnames gemaakt met BB FlashBack. Met een Tobii eye tracker werd nagegaan hoe de vertalers hun aandacht verdeelden tussen de verschillende onderdelen op het scherm. De definitieve vertalingen werden beoordeeld op hun kwaliteit door twee professionele reviewers, die een foutscoresysteem gebruikten. Deelnemers werden ook geïnterviewd zodat ze hun mening konden delen over bepaalde aspecten van hun prestatie.

De kwantitatieve data werden geanalyseerd met gemengde effecten lineaire regressiemodellen. De resultaten tonen aan dat vertaaltijd en typinspanning beïnvloed worden door vertaalmetadata, en dat de effecten variëren naargelang het type vertaalsuggestie (exacte matches, fuzzy matches, machinevertaling). Het onderzoek stelde ook vast dat er geen significante correlatie is tussen de prestaties van de vertalers wanneer ze typen, en de prestaties van de vertalers wanneer ze vertalen.

De kwalitatieve interviewdata tonen aan dat de manier waarop de vertalers hun eigen prestatie percipieerden en hun gemeten prestatie niet overeen komen. De vertalers verkiezen meestal de omgeving met vertaalsuggesties en metadata, zelfs wanneer deze omgeving niet gepaard gaat met betere prestaties. Vertalers vonden metadata een handige functie in de vertaalsoftware, onder andere omdat de metadata het makkelijker maken om hun vertaalstrategieën aan te passen aan het type suggestie. Vertrouwdheid met de taak werd ook aangegeven als een belangrijke factor die de percepties van de vertalers beïnvloedt.

De resultaten van deze studie wijzen op de noodzaak voor verder onderzoek over de interactie tussen vertalers en vertaalsoftware, met als doel niet alleen het bevorderen van productiviteit, maar ook van werktevredenheid. Ik hoop dat dit proefschrift ook bijdraagt aan de methodologie voor onderzoek over de werkplek, door een aantal uitdagingen en oplossingen te bespreken. Een belangrijke les is dat het noodzakelijk is om een optimale balans te vinden tussen ecologische validiteit en geldigheid van de data wanneer vertaalexperimenten worden uitgevoerd in realistische scenario's.

Acknowledgments

I would like to acknowledge the funding to my doctoral research, provided through the European Commission’s TIME Marie Curie fellowship (FP7-PEOPLE-2010-ITN-263954). It allowed me to dedicate full time to this research during three years, and provided all the necessary resources to purchase equipment, attend conferences and receive research training.

I would like to thank my supervisor, Anthony Pym, for his continuous support during my master’s and doctoral studies and for his insightful comments along the way. Thanks for never letting me lose focus on the “big things” and for helping me fight perfectionism!

I would also like to thank Reine Meylaerts, my co-supervisor and “promoter” at KU Leuven, for reading the thesis from a different perspective and contributing her invaluable knowledge and research experience.

I was very fortunate to have had the support of all the other participants in the TIME project: professors Christina Schäffner and Yves Gambier, and fellows Gabriel González Núñez, Wine Tesseur, Sara Ramos Pinto and Marta Miquel Iriarte. Thank you for generously sharing your friendship and experience during our workshops and project meetings.

I would also like to express my sincere gratitude to all the staff at MSS, for their generosity in providing all the information and all the support I needed for conducting my main experiment. In particular, I am deeply thankful to Raida Canut and Jordi Serratosa for all the support they offered during the four months I worked at their premises, and beyond. I would also like to thank Álvaro Rocabayera for opening the doors of his company and agreeing to become a partner in the TIME project. And I should not forget to give my sincere thanks to the translators who kindly agreed to take part in my experiments, an invaluable asset in any piece of empirical research.

Heartfelt thanks to Sharon O’Brien, for accepting to have me as a research visitor in DCU in the Spring of 2013, for her invaluable advice during a critical phase of my research and for her invitations to submit papers to conferences and books. Thanks to all the other colleagues I met in DCU, who made my stay a very pleasant one, with special mention to Joss Moorkens, Pat Cadwell and Sheila Castilho.

I am also indebted to Fred Hollowood, Katrin Drescher and Rafael Guzmán from Symantec Ireland, for contributing their business-oriented view of translation workflows in a big software company and for their kind hospitality.

A general thanks to all the great scholars I have had the honour to learn from during the research seminars I attended at the URV, in lectures and tutorials at CETRA research summer school, and in the many conferences I have attended, in particular to Andrew Chesterman, Arnt Lykke Jakobsen and Franz Pöchhacker.

I should also acknowledge the invaluable advice I received from the statisticians who helped me choose and apply the most appropriate models of statistical analysis, especially Oliver Valero Coppin and Laura Winther Balling.

A special mention go also to my office mates Esmaeil Haddadian Moghaddam, David Orrego Carmona, Esther Torres Simón, Nune Ayvazyan, Andrea Bellot and Kasia Baran, some of whom have already defended their theses and are doctors with honours. Thank you, David, for revising so many versions of my manuscripts and for always being willing to help me organise my ideas. Thank you, Esther, for helping me meet the deadlines, for your practical suggestions and for taking care of much of the paperwork. Another special mention goes to Alberto Fuertes Puerta, whose frequent incursions into the office were often enlightening. Thank you all for your friendship!

Aos meus pais, que direta ou indiretamente, contribuíram e contribuem para a minha formação como indivíduo e profissional. Ao meu irmão, com muito carinho.

À Marina, minha companheira, cujo apoio incondicional aos meus projetos de vida me permitiu retomar a minha carreira acadêmica e voltar a pensar nas minhas inquietações intelectuais, por acreditar no meu potencial. Graças ao seu apoio, terminei! Aos meus filhos, Gabriela e Nícolas, que são a grande alegria da minha vida.

Despite all the support I have received, any errors remaining in this thesis are my own.

Declaration

I, Carlos da Silva Cardoso Teixeira, hereby declare that this thesis is entirely my own work, carried out as a double doctorate between the Universitat Rovira i Virgili and KU Leuven, and that it has not been submitted as an exercise for a degree at any other university. Some parts of this thesis have been published previously in:

Teixeira, Carlos S. C. 2014. “Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories.” In: O’Brien, Sharon; Simard, Michel & Specia, Lucia (eds.) *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*. 45–59.

Teixeira, Carlos S. C. 2014. “Data collection methods for researching the interaction between translators and translation tools – An ‘ecological’ approach.” In: Schwieter, John & Ferreira, Aline (eds.) *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*. Newcastle upon Tyne: Cambridge Scholars Publishing. 269–286

Teixeira, Carlos S. C. 2014. “The handling of translation metadata in translation tools.” In: O’Brien, Sharon; Balling, Laura; Carl, Michael; Simard, Michel & Specia, Lucia (eds.). *Post-editing of Machine Translation: Processes and Applications.* Newcastle upon Tyne: Cambridge Scholars Publishing. 109–125.

Teixeira, Carlos S. C. 2013. “Multilingual systems, translation technology and their impact on the translator’s profession.” In: Neustein, Amy & Markowitz, Judith A. (eds.) *Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems*. Heidelberg and New York: Springer Verlag. 299–314.

Teixeira, Carlos S. C. 2011. “Knowledge of provenance and its effects on translation performance in an integrated TM/MT environment.” In: Proceedings of the 8th International NLPCS Workshop - Special theme: Human-Machine Interaction in Translation. *Copenhagen Studies in Language* 41. Frederiksberg: Samfundslitteratur. 107–118.

Tarragona, 9 January 2015

Carlos S.C T

Carlos da Silva Cardoso Teixeira

“What we gain in efficiency, we risk losing in humanism.”

(Pym 2004: 165)

Table of contents

List of tables

List of figures

List of abbreviations

BBF	BB FlashBack
CAT	Computer-Aided Translation
EBMT	Example-Based Machine Translation
GUI	Graphical User Interface
HCI	Human-Computer Interaction
LSP	Language Service Provider
NLP	Natural Language Processing
MT	Machine Translation
PE	Post-Editing
QA	Quality Assurance
RBMT	Rule-Based Machine Translation
RTA	Retrospective Think-Aloud
SMT	Statistical Machine Translation
TAP	Think-Aloud Protocol
TM	Translation Memory
TM/2	IBM TranslationManager
TPR	Translation Process Research
TS	Translation Studies
WYSIWYG	What You See Is What You Get

Types of translation suggestions:

E	Exact Match
H	High Fuzzy Match (85-99%)
L	Low Fuzzy Match (70-84%)
M	Machine Translation

Chapter 1. Introduction

1.1. Motivation

I started my career as a translator back in 1998. My first real translation project, for the Brazilian subsidiary of a French engineering company, had to be completed in about two weeks. The technical manual I was given to translate was very repetitive and contained a great deal of terminology, so I decided to use a translation memory tool for the task. That was a new technology for me, which I had just got to know a couple of months before, at an international translation conference I had attended at the University of São Paulo. Based on the comments on the translators' forum I used to follow at that time, I decided to try DéjàVu. The trial version of the tool was fully functional for a month, so I had enough time to learn how to use it and complete the project before the licence expired (and before the deadline!). To make the story short, I managed to complete the translation successfully, and that positive first experience made me an intensive user of computer-aided translation in my daily work as a freelance translator from then on.

Like most professional translators in technical fields, in the following years I learnt how to use several different translation tools, sometimes out of personal curiosity, sometimes due to my customers' requirements. Some tools were faster, others had more resources, and others were free. Some were more efficient for different types of projects, some worked better with the types of source files or translation memories provided by the customers. What all those tools had in common was the idea that professional translation should be based on translation memories, i. e. on previous translations done by human translators. For many years, I used to hear – and to repeat – that translation tools for professional translators had nothing to do with machine translation: "one thing is translation memory (TM), and a totally different thing is machine translation (MT)".

I maintained that dichotomous view until around 2009, when Trados offered the possibility to integrate machine translation into their translation memory system through a plug-in, and I decided to try it. The same tool manufacturer released a new product version that same year with built-in MT integration: when choosing the translation memories for a given project, now it was also possible to add a machine translation engine as a source of translation suggestions. The same approach was adopted by several other manufacturers around the same time and in the following years. The MT functions available had either restricted access (which would be granted to translators working on a particular project) or open access, by connecting to freely available MT engines. I believe the latter was mainly responsible for the gradual introduction of machine translation in the workflow of professional translators, as those free services allowed us to use MT alongside TM at no added cost. That happened at a time when machine translation output started to offer a much higher quality level than had been seen ever before, thanks to the development of statistical approaches.

With this workflow, when no “good” option was available from the translation memory, I could type on top of machine-translated text instead of translating a segment from scratch or by editing the source text. New possibilities arose as a consequence of this new way of working, as well as many questions. The spreadsheet where I used to keep track of all my translation projects was indicating a productivity gain of up to 30 percent, but that seemed to depend on many factors. The industry standard of only retrieving fuzzy matches above the 70-percent level – as TM matches below that level were presumed to be more time-consuming or error-inducing than translating from scratch – started to be challenged. Now, with the possibility of working from a machine translation suggestion, was it still a good idea to retrieve matches as low as 70 percent? What about the order of presentation of suggestions: should TM matches of any level be always presented first, before MT feeds? And what about the useful information provided about TM matches – which throughout this thesis I am calling *translation metadata*: how would the lack of equivalent information affect the handling of MT suggestions?

When I decided to go back to university and do research on translation technology, those were the topics and questions that occupied my mind. By reading the literature in the field, initially through the works of Sharon O’Brien and Arnt Lykke Jakobsen, I realised the possibilities for exploring those questions, in terms not only of topics but also of the many tools that made it possible for researchers to study the translation process. At the same time, I realised the literature on translation technology in general and on the combination of machine translation and translation memory in particular was rather scant.

It was in this combined context of personal curiosity based on my own professional experience, on the many possibilities offered by the current research tools and on the several research topics that remained to be explored that I came up with the idea for my doctoral research (preceded by a research masters where I started to develop the methodology). I decided to investigate what happens if we eliminate all translation metadata from the translation memory environment and work in a way more similar to pure MT post-editing. The idea was to create a task to be translated in a traditional TM system, with translation metadata, and a similar task to be translated in the same system, but without metadata. Then both tasks would be compared in terms of productivity, effort and quality. The present thesis is the result of that research endeavour.

1.2. Aims and objectives

The research question that this thesis seeks to answer can be stated as follows: What are the differences (if any) in the translation process between a situation where translators have access to the metadata about the translation suggestions and a situation where the metadata are not available?

A deeper understanding of how translators process the information available on screen can help improve their workflows and practices, and it can also help enhance the ergonomics of translation tools. Those improvements can bring benefits for all parties involved in translation projects, including translators, translation agencies, translation-tool developers and, ultimately, translation customers.

Finding optimal processes and tools can also increase the volume of text that can be processed. In the European Union, as in many international organisations, large amounts of text remain untranslated due to time or budget constraints. Higher efficiency can help reduce those limitations without adverse effects on customers' expenses or translators' earnings. This can lead to greater dissemination of information, wider access to foreign markets by companies seeking to sell their products and services, broader access to legislation in national languages, and the empowerment of speakers of minority languages.

As a second goal, I am also trying to understand the cognitive and emotional aspects involved in the translation process. For example, despite any differences in productivity, in which environment do translators *prefer* to work, e.g. does the presence of metadata make them feel more comfortable? In a broader sense, I hope to obtain results that are of intellectual importance, by understanding how technology can affect the decisions made by translators when translating, not just in terms of efficiency but also with respect to the affective aspects of job satisfaction. Finally, I wanted to investigate up to what extent the translators recognise metadata as indicators that (some) translation suggestions have a slightly humanised provenance.

1.3. Overall research design

This thesis focuses on a particular aspect that distinguishes translation memory (TM) systems from machine translation (MT) post-editing environments: TM systems show translators the metadata (origin, author, textual differences, etc.) of the translation suggestions coming from the memory, whereas most environments for post-editing MT display the best translation suggestion possible, without any metadata. The presence or absence of translation metadata might influence translators' performances and perceptions, but very little research in the field has focused on this particular distinction.

In order to address the topic, I compared two translation tasks. In one of them, the translators can see the metadata on translation suggestions (Visual task), whereas in the other task they do not have access to this information (Blind task). One part of my study consists of testing whether the availability of metadata affects translation time, typing effort and error scores, and whether any effects depend on the type of translation suggestion. Another part of the study includes an investigation of how the translators *perceive* the two tasks.

The study is based on a translation process experiment with 10 professional translators working from English into Spanish. Each translator was asked to perform the two main translation tasks mentioned above: one task in a Visual environment (with translation metadata) and one task in a Blind environment (without translation metadata). The order of tasks and source texts was evenly distributed among the participants.

The performances of each translator were assessed with process-research tools that include keystroke logging, screen recording and eye tracking, as well as with human quality evaluation and interviews. The quantitative data were analysed statistically using mixed-effects regression models, in order to test for the effects of translation metadata and type of translation suggestion on the three main dependent variables, i.e. translation time, typing effort and error score. An exploratory analysis was also done by including additional predictors in the statistical models. Finally, I explore the data obtained in the interviews to assess the translators' perceptions, and I analyse the screen recordings and eye-tracking data to tap into the translators' actual behaviours and strategies in specific segments.

The two source texts to be translated in the two main tasks contained around 500 words and were extracted from an IBM software manual in such a way as to produce two texts with 28 segments each. Two translation memories were created (one for each of the source texts) from legacy customer memories and machine translation (a customised Moses engine), so that the texts were presented with four types of translation suggestions (seven segments of each type): exact matches, fuzzy matches of 70-84%, fuzzy matches of 85-99% and machine translations.

IBM TranslationManager was used as the translation tool within which the two translation environments were reproduced. Inputlog (Leijten and van Waes 2013) was used as the main keystroke-logging tool to measure the amount of time and the number of keystrokes used by the translators to produce a given translated segment. BB FlashBack was used as a screen-recording and face-recording tool to indicate what the translators were doing at any given moment. By recording and then watching the activity of each participant translator, I was able to view how they dealt with each particular segment. A Tobii X120 eye tracker and the Tobii Studio software were used to identify where on the screen a translator was looking at during a translation task.

1.4. Translation metadata

Metadata can be generally defined as “data about data” (Anastasiou and Morado Vázquez 2010: 257) and can come in many forms depending on their use and application. “Translation metadata”, as I define the term, is the information that appears on the interface of a translation tool to inform the user about several aspects of a translation task, in addition to the source text.

In general, translation metadata can include a vast range of elements, such as the language pairs involved in a project, translation progress statistics, the state of segments (translated, not translated, automatically propagated, reviewed, pending, approved), terminology assistance from term bases (glossaries) and information about translation suggestions. This last set of metadata elements – information about translation suggestions – is the focus of this thesis. It can be divided into two broad categories: provenance metadata (indicates whether a translation suggestion comes from machine translation or from a translation memory) and translation-memory metadata. When a translation suggestion comes from machine translation, no further metadata are displayed; the remaining metadata elements displayed by the tools concern translation-memory matches. An exhaustive list of metadata elements and a tentative categorisation of those elements are provided in Appendix 11. For the sake of simplification, “translation metadata" is used elsewhere in this thesis to refer only to metadata about translation suggestions.

It has been argued that the presence of translation metadata is a typical feature of translation memory systems (Anastasiou and Morado Vázquez 2010; Karamanis et al. 2011; Morado Vázquez 2012; Teixeira 2014b) that helps translators to make choices among different types of suggestions. This thesis tries to contribute to the debate by analysing the effect of such elements on translators' performance.

1.5. Thesis structure

Chapter 2 gives an overview of what can be understood by translation technologies, covering translation memory, machine translation and the integration of both. In Chapter 3, I present a brief review of the literature on translation technology, focusing on the main publications in the field that have contributed to the current study.

Chapter 4 presents the Methodology. After introducing the research question, the hypotheses and the operationalisation of variables, the chapter describes a pilot study that was done in preparation for the main experiment, with some results and lessons learnt. Then the chapter presents the main experiment: it first explains the context in which the experiment was carried out, then presents the participants and materials, and finally describes how the experiment was actually run. Another section within this chapter explains the data collection methods, which include keystroke logging, screen recording, eye tracking, interviews and translation reviews. The chapter goes on to describe the equipment and software used for the data collection and explains how the data was analysed. It concludes by mentioning the ethical concerns that were taken into account when preparing and running the experiment, and later when dealing with the collected data.

Chapter 5 presents and analyses the Results of the experiment. It starts by presenting the general quantitative results, where simple descriptive statistics is used to look at the distributions and to find correlations between the two main translation tasks and two preliminary tasks. It then presents the qualitative data collected in the interviews and relates the participants' perceptions with the quantitative data presented in the first section. Section 5.4 presents the bulk of the data analysis, where the quantitative data for the two main tasks are analysed in detail using inferential statistics, through mixed-effects models. Finally, the last section in this chapter presents a brief analysis of the eye-tracking data, taking some “rich points” (PACTE 2005: 614) in the translation as examples. The additional information provided by the translators’ gaze behaviour is used for analysing how they tackled the translation tasks under the two conditions presented in the experiment (with and without translation metadata).

Chapter 6 discusses the findings. It starts by testing my three hypotheses and sub-hypotheses, then discusses additional findings and relates those results with the information collected in the interviews.

Chapter 7 contains my conclusions. It summarises the findings, discusses their potential applicability and suggests some contributions the thesis might have made to the field. It also lists several shortcomings and limitations of the present study and presents avenues for future research.

Chapter 2. Overview of current translation technologies

Translation technologies encompass a wide range of tools that can help to produce translations. Nowadays, those technologies appear predominantly in electronic format, but they have been present over the centuries in simpler formats, such as paper dictionaries, notebooks and typewriters.

Contemporary translation technologies are usually referred to under the umbrella name of Computer-Aided Translation (CAT). The most prominent form of CAT tool is the "translator's workstation", a concept envisaged in the late 1970s and early 1980s by people like Martin Kay (1980) and Alan K. Melby (1982), and still in full effect today. The workstation, or workbench, consists of an integrated, computer-based environment for translating electronic files, and puts the translator in the centre of the process. The first CAT tools appeared in the form of terminology management systems in the early 1980s and then as translation memory (TM) systems in the early 1990s (Somers 2003; Christensen and Schjoldager 2010; Melby 2013). CAT tools have evolved to also include several other functions, such as project management and quality assurance, and have extended their scope to include machine translation. A recent and comprehensive compilation of existing CAT tools can be found in Zetzsche (2014).

Some criticism has been expressed of the term "CAT tool". Pym (2010) says "[t]he term is misleading, since almost all translating is done with computers these days, so all processes are 'computer-aided' to some extent" (Pym 2010: 123), or again, "[t]he term [...] is now a misnomer, since computers are involved in almost all translations jobs, and in a lot of interpreting as well" (Pym 2011a: 78). In defence of the term, one should say that it mimics similar terms used for other fields, such as Computer-Aided Design (CAD), Computer-Aided Manufacturing (CAM) and Computer-Aided Engineering (CAE), and one can hardly imagine the activities in such technical fields to be performed without the help of computers either. Yet those terms have consolidated over the years (even though the tools themselves have changed a lot – Autodesk's Autocad being a paradigmatic example) and refer to a known set of tools within the profession.

One should agree with Pym's criticism when he goes on to suggest that "[t]he term should be replaced by clear reference to the technologies actually involved (e.g. translation memories, machine translation, terminology database)" (Pym 2011a: 78). The same point is brought up by Zetzsche (2014: 189):

> CAT, or computer-assisted translation tools, is a great term for describing the numerous families of software tools that translators use for their work [...]. Unfortunately, we often use “CAT” as a synonym for so-called “translation memory tools”, when the latter is really only a sub-category of the former.

Zetzsche advocates the use of the term “Translation Environment Tool” (TEnT) instead, which he considers “describes much more accurately the various ways that we should use these tools in our translation work” (Zetzsche: loc. cit.). However, his term has not found an echo among translators and translation scholars, who continue to use CAT tools (and its specific sub-components). So the solution I am using here is to talk about “CAT tools” in general, or to specify the particular sub-components, especially “translation memory systems”, when necessary.

Another set of translation technologies that deserve specific attention can be grouped around machine translation (MT). In contrast with Melby’s vision of a translator-centred workstation, the first approaches to machine translation sought to build a tool that would be able to translate automatically (independently of human intervention). Instead of having the translator at the centre, early MT implementations were based on the belief that it would eventually be possible to replace human translators entirely. Failure to achieve such an ambitious goal changed the direction of MT research over the decades. The prevailing view nowadays is that MT makes more sense as a tool to *help* the translation process, and the currently relevant debate concerns how best to integrate it into the translation workflow and into CAT tools. One of the most common uses of machine translation nowadays is *post-editing*, a process in which a translator or post-editor receives the initial output from the MT engine and repairs (edits) it if necessary to produce the final translation. Other strategies such as pre-editing and controlled language are also used in order to improve the results. More recently, new forms of integration of MT into the translation workflow are being explored with successful results.

In the following sections, I will cover the basics of translation memories and machine translation, and introduce other concepts that are frequently mentioned in the literature and used throughout this thesis.

2.1. Translation memory

A translation memory (TM) is a database that contains chunks of source texts paired with one or more translations for that source text, plus associated metadata (information on the pair). A translation memory system or translation memory tool is a computer programme that is able to store those "records" of source text / target text / metadata and to later retrieve useful information from the database for the translator, based on a set of "matching" rules. The propagated benefits of TM systems include their potential to increase productivity, improve terminology consistency and reduce repetitive tasks.

One common translation memory workflow consists of the following steps: the translator configures the tool with a translation memory and then opens a source file for translation. Based on the file format, the tool segments the text to be translated, usually according to standard punctuation marks such as periods and colons. The source text then becomes a sequence of source segments, which are presented sequentially for translation. When the translator "opens" a segment, the tool automatically searches the database (the translation memory) to check whether there are similar source segments stored there. If there is a segment in the database that has exactly the same source text as the source text in the active segment, we say that an "exact match" has been found. If there is a segment in the memory whose source text is approximately the same as the active source text, then we talk about a "fuzzy match", which has an estimated similarity level associated with it. When no similar text is found in the database, then we say there is a "no match". For each source segment found in the database, the tool presents the corresponding target segment(s) and some metadata, which may indicate the type of match (exact or fuzzy), the similarity level, the author of that translation, the date it was produced, etc. (see section 1.4 and Appendix 11 for a more detailed explanation of translation metadata). The translator can then choose from the various suggestions retrieved by the tool and produce the final translation based on the chosen suggestion. Once the translation for a segment is finished, the translator moves to another segment and the complete translation pair is saved in the translation memory for future use.

This basic description of the translation memory workflow applies to virtually all translation memory systems available in the market. The differences lie in several details that are specific to each tool and configuration; for example, the file types the tool is able to import, the segmentation rules, the algorithm for retrieving matches from the memory[1], the amount of metadata it can retrieve from and store in the database, the layout of the text presentation on screen, the keyboard shortcuts available for carrying out the different functions in the tool, the way of integrating terminology management and quality assurance, and the possibilities for integrating machine translation into the workflow.

Those differences between the translation tools make them more suitable for different domains (e.g. software localisation vs. technical documentation) or for different translating styles. Some tools are better for collaborative workflows, as they offer the possibility to share the same translation memory among several translators by storing it on a shared server or “in the cloud”[2]; other tools are web-based instead of requiring installation on a local computer; and so on.

The possibilities offered by TM technology of quickly retrieving previously translated text have increased the amount of text that can be processed using the same (human) resources. For example, translation memories have made it easier to update large documents more frequently: instead of having technical writers indicate the new portions of text that require translation (or comparing older and newer versions of the same document), TM tools can automatically indicate what has changed, speeding up the translation process and making sure that consistency is maintained with the previous versions of the document (at least in theory).

These possibilities have in turn allowed for new text-production strategies, which involve non-linear ways of producing texts as well as non-linear ways of translating texts (see Pym 2004: 185–188, on the “loss of discursive linearity” in localisation), satisfying the demands posed by the global distribution of content. Newly produced texts are non-linear because they are “leveraged” from existing texts, and the process of translating those texts is also non-linear, not only because the source texts are fragmented, but also because translation technology “leverages” previous translations as well.

Between the late 1990s and the middle of the following decade, the use of translation memories became standard practice in several specialised domains, first in software localisation, then in technical, medical and some types of legal translation. The

[1] Carl and Hansen (1999) remind us that the algorithms behind TM matching have the same origin as those used in machine translation, especially EBMT.

[2] The “cloud” refers to the distributed storage and processing of information on different servers on the Internet.

next technological leap would come later in the decade, when the use of machine translation started to be introduced at large scale in the language industry.

2.2. Machine translation

Although the idea of automatic translation can be traced to as early as the 17th century, machine translation in its contemporary form dates back to the first half of the 20th century (Hutchins and Somers 1992; Somers 2003). It has undergone several stages of development, with the current trend being towards data-driven statistical machine translation (SMT) and hybrid approaches, which combine SMT with rules-based MT (Way 2009).

In general, a machine translation engine receives the source text in the source language as input and produces the target text in the target language as its output. The method for converting the source text into the target text varies depending on several factors, such as the type of engine and how it is configured. Rule-based machine translation (RBMT) uses complex linguistic rules (morphological, syntactic and semantic) and a large volume of bilingual dictionaries to translate texts from one language to the other. Such systems can be built for any language combination, as long as the rules and dictionaries are defined by expert linguists. This usually requires a process that is costly in terms of both time and money. RBMT can produce translations of good quality, depending on the domain (the more structured and unambiguous the sentences, the better the results) and language combinations (the more closely related the two languages, the better the results). Once such a system is built, the output tends to be consistent, but difficult to improve further. Despite the proposal of example-based machine translation (EBMT) by Nagao (1984), RBMT systems dominated the MT research agenda until the late 1980s, when the radically new paradigm of statistical machine translation started to develop in IBM research centres (Brown et al. 1988).

Current SMT systems are built from large corpora of bilingual texts and rely on computational power to build statistical translation models. They have the ability to “learn” by training, so the larger the training corpora and the better its training rules, the more accurate an SMT system will be. Given a source text, the engine applies statistical rules of probability to find the best match in the target language. The quality of the output depends less on the similarity between the two languages than on the availability of training corpora. Thus, this approach is more useful for those language combinations that have enough material to train the system, where "enough" in this case can mean several millions of words. Building an SMT system requires great computational power, but this is becoming less of a problem with increasingly higher computer capacities and also thanks to cloud computing. SMT is more suitable than RBMT for dealing with uncontrolled source texts (such as user-generated content) and tends to deliver translations of higher quality in terms of fluency. The developments made by Google in the late 2000s constituted a great quality leap for SMT in particular and for machine translation in general, with its Google Translate engine being made freely available in the public domain.

SMT has some shortcomings too, such as a difficulty to keep terminological consistency. This can be compensated for by tuning the baseline system with client-specific corpora, resulting in a customised engine whose output is more in accordance with the company's style and terminology for a given domain. Another strategy for improving consistency and also for handling *tags*[3] is to complement the statistical model with a set of linguistic or post-hoc rules (e.g. regular expressions). These systems are called "hybrid" and have become the general trend in recent years. For a synthetic historical overview of the different MT paradigms, see Way (2009).

Some aspects have prevented or slowed down the adoption of machine translation in the translation industry. The first issues are related to the MT technology itself. For example, the best-performing freely available engines are of a generic nature (not customisable), which means that they do not provide enough quality for specialised texts, especially as far as terminology is concerned. Customised engines are expensive to implement and maintain, and require large volumes of training data. In practice, language-service providers (LSPs) need to hire specialised people to interact with the MT system or pay an external company to house and maintain their system. Whether or not the investment will pay off depends on the volume that the company can translate with machine translation and the savings resulting from the use of this technology. Moreover, most of the free online engines do not ensure the privacy of the contents submitted, generating concerns related to intellectual property and industrial espionage. The second

[3] In markup programming languages such as HTML and XML, a tag is part of a structure that contains instructions on how the file content should be processed. In a typical translation-memory file (whose main standard is TMX, based on XML), tags serve as delimiters for style and content. Tags are not part of the translatable text and can cause difficulties to MT systems, such as: (1) the need to interpret the source segment correctly, with some breaks in the normal syntactic flow; and (2) the need to reproduce the segment in the target language with the tags in the right places.

set of issues is related to the human factor. This includes aversion to changes in established work methods and to the implementation of new technologies in general, not only among translators but also among project managers and company owners (see Pym 2011b: 5 on the resistance of professionals against changes in technology). The revision of machine-translated text (post-editing) has not been particularly well received by translators, perhaps because they do not like revision tasks in general, or because the errors that are present in machine-translated text are of a more “stupid” nature (see O’Brien and Moorkens 2014), or because translators feel that machine translation limits their creativity and the richness of the target text. Despite all those hindrances, machine translation has gained terrain progressively and industry estimates for the coming years tend to be optimistic, even predicting that “[p]ost-editing MT output is likely to overtake translation memory leveraging as the primary production environment in industrial translation in the next five years” (van der Meer: 7).

In recent years, research and development in the field of machine translation has definitely started to focus on finding innovative ways for human translators to use the output of MT engines in production settings. For example, the CASMACAT project (CASMACAT 2014) has introduced the concept of Interactive Translation Prediction (ITP), a translation mode in which the workbench populates the target segment with a suggestion from the MT engine and, as the translator starts repairing the suggestion, the system adapts the remaining of the suggestion accordingly, in a continuous process. The CASMACAT project has also been developing a post-editing environment that can display confidence estimates as metadata for machine translation suggestions, along the same lines as what has been done by the PET project (Aziz et al. 2012). Finally, the CASMACAT project has introduced a translation mode that learns from the human translator’s edits and retrains the MT engine continuously, a strategy that has also been offered by commercial tools such as Sovee.

While no machine translation system has achieved a quality level comparable to high-quality human translation, many of them have achieved acceptable quality for different purposes. Outside of the world of translation professionals, for example, it is not uncommon for users to rely on their web browser’s automatic translations when viewing an Internet page in a different language. Although the translations are normally not perfect, they usually fit the purpose of giving a general idea of what the page is about (gist translation), even allowing the user to make decisions or perform some actions, like purchasing a service or product.

This greater accessibility to translations by direct users and non-professionals is another reason why, in the translators' trench, there has long been apprehension and suspicion with relation to MT. A survey conducted by Piróth (2011) with 160 translators from the IAPTI, ATA and ProZ.com actually reflects the mixed feelings regarding the use of MT among translators (see also Hartmann 2010). This happens especially when MT is seen as a potential replacement for human translators and when the great potential of MT in helping the work of professional translators is disregarded.

Similarly to what happened when translation memories were the "new technology", one can expect that machine translation and future translation technologies will allow more texts to be translated, actually increasing the available workload for translators – at least, for those who are able and willing to use those technologies.

2.3. TM & MT integration

Since MT has started to be introduced more systematically in the realm of professional translation, several strategies have been tried to integrate the new technology into existing workflows, which were centred on the concept of translation memories. One initial approach was to use MT post-editing for certain tasks (depending on the domain, file type, client, etc.) and continue with a TM-only approach for the remaining tasks. However, a logical next step was to find ways to combine both technologies in the same workflow, which is still the general tendency at the time of writing this thesis.

In a typical workflow involving TM and MT among LSPs, the source files to be sent out for translation are first processed through the available translation memories to isolate any segments that would yield a "no match". These segments are then processed through a machine translation engine, which provides a suggested translation for each segment, in the form of a new "translation memory". The project manager prepares the project combining one or more translation memories from previous projects with the "translation memory" generated by the machine translation engine. When the translators receive the file for translation they thus get at least one suggestion per segment (what would have been a "no match" in a typical TM-only scenario now receives a machine translation suggestion).[4]

[4] This workflow has been reported as being in use as early as the late 1990s (see Webb 1998: 53).

Another typical workflow, more common among freelance translators, is to have the "no matches" replaced dynamically as they translate: for each segment, the translation tool will retrieve suggestions not only from the available translation memories but also from the available (on-line) machine translation engines.

In both workflows, the translation cum post-editing is normally done within traditional translation memory systems, which indicates a convergence between both technologies. Another sign of convergence is that the statistical MT engines are based on the use of large databases of previous translations, which are composed mainly of translation memories. Moreover, TM systems such as DéjàVu offer the possibility of sub-segment matching, providing a hybrid TM/MT suggestion assembled from chunks of the TM using MT algorithms.

One issue that has arisen in the translation industry since machine translation was introduced at production scale is how to compensate translators for their post-editing effort. The payment schemes for translation memories are already well established, with discounts applied according to the match types (the "Trados table"). However, machine translation suggestions do not offer a precise "fuzzy match" number to base any discounts on. The solution used by some in the industry has been to apply post-hoc discounts, based on how much editing effort was actually invested in repairing the MT suggestions. For example, MemSource markets a family of translation tools that offer a way of calculating a rate of pay for post-editing based on post-task calculations. It tries to mimic the way other TM systems calculate the rates of pay for fuzzy matches and applies the same kind of grid for post-edited machine translation proposals. As the tool manufacturers themselves recognise, this approach requires a change in the way translation projects are invoiced, as customers are used to requesting quotations in advance. The MemSource approach requires a business relationship in which the translation client trusts (based on the tool reports, of course) the language service provider that the fair amount for the post-editing effort is being charged. Although this seems improbable at first sight, major translation customers handling millions of words a year might very well be ready to work this way (in fact, IBM has already been working this way for some years, relying on the log files generated by its own CAT tool, IBM TranslationManager).

Chapter 3. Previous studies on translation technology

O'Hagan (2013: 505) mentions that new technologies have affected translation practice in two different ways: first, at the "micro level", in the form of electronic tools used by translators; second, at the "macro level", by affecting the type of content that is translated and also by promoting new ways of collaboration and participation, as in crowdsourcing initiatives. O'Hagan adds that new technologies have in turn contributed to translation research by providing tools that help investigate the translation process as it unfolds.

Historically, the foci of research on translation technologies within Translation Studies have evolved in parallel with the evolution of the technologies themselves, although there has always been a natural lag between the release of new technologies and the publication of research papers about them. Tool manufacturers and the translation industry in general have published extensively on the different types of tools, while translation scholars have tried to keep up with the developments in a more conceptual way (cf. Pym 2012 for the different kinds of contributions that can be expected from either community). Researchers in the fields of Computational Linguistics and Natural Language Processing (NLP) have published on the more technical aspects of the technologies.

In this chapter I will give a short overview of studies on translation technology, from different perspectives. First I will focus on studies that deal with how translation is performed with the help of translation technologies. Studies on the two main technologies covered in the previous chapter (translation memory and machine translation) will be mentioned, as well as studies that have looked at the interaction of both. This will introduce the concept of translation metadata, which is the central topic of this thesis. Then I will also refer to studies that focus on translation technologies from the perspective of human-computer interaction, taking into account factors such as ergonomics and usability. This chapter will conclude by suggesting what is still lacking in the existing research and by contextualising the reasons why I decided to focus on translation metadata as my main research topic.

3.1. Translation memory

After the first commercial TM systems came into existence in the early 1990s, translation researchers started to publish on the topic later in the same decade. Webb (1998) offers a detailed definition of TM databases and an explanation of how the functions of analysis, concordance and matching work. Based on surveys and case studies, she gives an overview of market relations between translation buyers, translation agencies and freelance translators at the end of the 1990s. Webb presents an analysis of productivity and quality gains based on factors such as type of project, availability of source files in electronic format, text type, availability of previous reference material and the frequency of updates. Her study finishes with a prophetic foresight for the following decade:

> The future looks bright for all kinds of CAT tools. In fact, the future of translation is heading toward automated solutions. This doesn't necessarily mean that machine translation will dominate the translation industry. While machine translation is improving, it is only part of the bigger picture. The trend is to integrate all of the different CAT tools and elements into one continuous process. (Webb 1998: 53)

That same year of 1998 saw the first edition of an important introductory text book, republished two years later (Esselink 2000). Esselink introduces the basic concepts behind translation memory systems, with examples of real tools and special mention to Trados Translator's Workbench, STAR Transit, Atril DéjàVu, SDLX and IBM TranslationManager. It is worth noting the list of "disadvantages of translation memory" provided by the author, such as the difficulty to "see how translated text will be displayed in the final layout", the sharing of translation memories among "several (teams of) translators in different locations", the difficulty of "keep[ing] the TM databases up-to-date" due to "last-minute changes [...] in the translated files", the lack of filters for specific file formats, and the impossibility of "chang[ing] the overall structure of the text, i.e. [of] chang[ing] the sequence of sentences within a paragraph" (Esselink 2000: 367). Some of those disadvantages have been partially addressed over the years, with the creation of additional import filters, increased connexion speeds that make it feasible for more than one translator to work on the same TM over the Internet, and preview panes that allow translators to visualise the target text in a WYSIWYG manner as they translate (see Biau Gil 2005 on the effects of the lack of visual context in TM tools). Other disadvantages mentioned by Esselink still apply today, such as the difficulty of keeping translation memories up to date and the impossibility of changing the order of sentences (at least without a compromise to the coherence of the translation memory), due to the paradigm of a fixed linear segmentation underlying most TM systems. This type of segmentation is usually taken for granted but can have an impact on all the metrics relevant for the field (see Dragsted 2004).

Austermühl (2001) mentions the benefits of TM usage as including an “increase in income”, the “elimination of repetitive translation tasks” and “consistency” (Austermühl 2001: 140). Like Esselink, he warns, however, that “‘mistranslations’ are also subject to repetition and reproduction” (loc. cit.), which should be compensated for by maintaining the TM database on a regular basis (on this topic, see Moorkens 2012).

One of the first empirical studies to deal with translation memories is Dragsted (2004). This doctoral thesis investigates an important aspect of translation memory tools: how the forced segmentation (usually sentence-based) in the tools relates with cognitive segmentation. Dragsted uses keystroke logging and retrospective verbalisations to collect data from professional translators and translation students. She finds that while participants in both groups tend to prefer segmentation at the paragraph level, they actually process units that are smaller than the sentence (especially when translating difficult texts). Based on this empirical evidence, Dragsted recommends that “TM systems be adjusted so that the focus is removed from the sentence, while at the same time segments below the sentence level are retrieved [from the translation memory]” (Dragsted 2004: 280). Not only does she challenge the established practice of sentence-based segmentation in TM systems but she also provides additional evidence of translators’ preferences regarding certain configurations in the tools. A more concise version of this study has also appeared as Dragsted (2005).

Colominas (2008) is another study that deals with TM segmentation. It shows that sub-sentential segmentation (at the noun-phrase level) can increase recall (matching) up to 25 percent, with various degrees of precision (usefulness) of the suggestions. The study was carried out on texts from the European Parliament and the United Nations, and in this regard stands out from most of the studies in the field, which tend to favour texts in the technical and localisation domains. The conclusion that sub-sentential segmentation has the potential to improve the leverage of translation memories is in line with the findings by Dragsted (2004; 2005) and calls for increased awareness of the possibilities of segmentation in the design of TM tools.

Wallis (2006) analyses the translation process under different workflows by comparing the performance of translators working with pre-translated text as opposed to interactive suggestions. Although she finds no significant differences for productivity and quality between the two modes, the translators’ satisfaction was greater in the interactive mode. Her study was done with only four participants (translation students) and included subjective data collection methods such as self-reporting of session times and post-performance questionnaires. Nevertheless, it is relevant for the methods it employs and for taking into account not only productivity metrics but also the opinions of the participants.

Christensen and Schjoldager (2010) review the available empirical research on TM published over the decade and call our attention to the low volume of studies on the topic, especially on aspects related to the interaction between translators and the technologies:

> Little research has been carried out on how translators interact with TM and how TM systems affect the cognitive (internal) translation process, and very few studies of TM are empirical investigations. (Christensen and Schjoldager 2010: 89–90)

Over the following years some other studies have appeared, most of them published as doctoral theses. Yamada (2011) investigates how the type of content (“free translation” vs. “literal translation”) in a translation memory affects translation speed, and concludes that literal translations are more advantageous for higher fuzzy-match categories. Martín-Mor (2011) studies the effects of TM systems on linguistic aspects of the final translation, according to the tool environment and different translator profiles. Moorkens (2012) shows how inconsistencies propagate in TMs and how much effort is necessary to maintain a clean TM (in a process called TM “laundering”).

However, the still scant number of publications on translation processes with TM seems to be reflecting two things: on the one hand, that the research methods available for translation process research are not solidly established yet or cannot be used in combination with TM systems (e.g. Translog); on the other, that TM might not be the main focus of interest, as other studies are being published on other aspects of translation technology, especially on MT post-editing.

3.2. Machine translation

The main contributions from the Translation Studies community in relation to machine translation regard the potential uses of MT and the types of interactions between the technology and the people using it.

According to the panorama presented in Austermühl (2001) of machine translation at the end of the 20th century, MT “architectures” were mostly rule-based, with commercial offers broadly divided into “low-end systems” and “high-end systems”. Low-end systems were for “non-translators or casual users” looking for “indicative [gist] translation” with low quality expectations, while high-end systems targeted big corporations and organisations, seeking to increase productivity for their large-volume translation projects in technical fields, with variable levels of quality expectations.

Esselink (2000) refers to MT in the same period as not having “been used extensively in the localisation industry”, although he foresees that the situation might “change in the near future” (op. cit.: 394). His description of MT technology is also limited to the rule-based paradigm and still reflects a clear-cut dichotomy between MT and TM. However, Esselink envisages the potential of MT “as a translation productivity tool, rather than a replacement for the translator” (loc. cit.).

Similarly, Austermühl does not present MT as being antagonistic to TM, but rather as having a major role in the production of translations: “MT is an aid to (not a replacement of) professional translators” (Austermühl 2001: 168). According to Austermühl, human intervention with machine translation could happen before (pre-editing), after (post-editing) or during the generation of MT output (interactive mode). In the last-mentioned, the system pauses “to consult the user when it encounters problems it cannot resolve [...], for example [...] syntactic or semantic ambiguities” (Austermühl 2001: 165). This mode seems to have been abandoned in later implementations of MT, in favour of newer approaches such the Interactive Translation Prediction (ITP) mode offered in MateCat (Federico et al. 2014) and CASMACAT (Underwood et al. 2014).

Pre-editing, on the other hand, continues to be used, mainly through controlled language (O’Brien 2010). This consists in following specific guidelines when writing the source text, in order to avoid complex syntactic structures and other textual issues such as ambiguity that are known to create difficulties for the MT engine. It can also include language checkers to make sure those guidelines are correctly followed (Bernth 2006).

Post-editing (PE), in turn, can be generally defined as “the task of editing pre-translated text that has been processed by an MT system from a source language into (a) target language(s)” (Allen 2005: 1) or “checking, proof-reading and revising translations carried out by any kind of translating automaton” (Gouadec 2007: 25).

Post-editing is usually classified in two main categories, according to the expected quality level, which in turn is based on the purpose intended for the final translation: “light post-editing”, when post-editors fix only the most severe errors that could cause misunderstandings, and “full post-editing”, when they bring the text to the highest quality, e.g. for publication purposes. Additional intermediary levels are also mentioned by Allen: "minimal PE, rapid PE, partial PE, maximum PE" (2005: 1).

Post-editing as a form of human interaction with MT output has been the focus of several studies such as Guerra Martínez (2003) and Mossop (2001; 2007). Post-editing has also many interfaces with revision processes in general (Allen 2003). The basic differences between post-editing MT and revising human translations lie in the types of errors that tend to be produced by MT engines and by human translators (cf. O'Brien 2002: 101) and perhaps also in the trust that post-editors or reviewers attribute to those different "authors" of the translation.

One of the most detailed studies on post-editing is still Krings (2001). Although the post-editors in the study worked on paper rather on a computer (due to technical limitations at the time when the experiment was conducted – in the early 1990s), Krings presents a sound analysis of the post-editing process in terms of mental operations. The study involves 52 participants and is based to a large degree on think-aloud protocols (TAPs), which are coded and classified. Krings proposes a model for post-editing effort with three components: temporal, cognitive and technical. Although new technologies have emerged since the publication of this study in terms of MT systems, post-editing tools and research tools, Krings (2001) remains an important reference for post-editing process research for the detailed analyses and considerations he presents.

Several studies have tried to determine whether there is an actual increase in speed or quality while post-editing machine translated segments when compared to translating from scratch. Allen (2003; 2005) has conducted a series of studies on MT post-editing with specific tools and provides some guidelines for improving its results. Many cases of increased speed are reported in the industry (e.g. Plitt and Masselot 2010; Skadiņš et al. 2011). These studies have found that trained MT systems that are used for translating restricted-domain texts can increase translation speeds significantly when compared to translating from scratch. On the other hand, Garcia (2010) compares time and quality between translating "entirely from the source text" and "editing machine translation" from a generic statistical engine (Google Translate). In this case, he finds that "time differences were not significant", although "the machine translation seeded passages were more favourably assessed" (op. cit.: 7). It is apparent from several studies that the gains in productivity actually depend on factors such as the quality of the MT engine and text type.

Lee and Liao (2011: 142) "suggest various benefits for the use of MT, such as facilitating source text comprehension and reducing translation errors." Nevertheless, the authors' productivity assumptions that MT can save much time "from needing to type out words" (op. cit.: 141), although plausible, require further empirical testing. A study with seven participants working in the English-to-Danish language combination (Carl et al. 2011) actually points in the opposite direction, as will the present thesis.

De Almeida (2013) focuses on how factors such as experience affect the post-editing process. She uses a mixed-method approach that combines quantitative data from keystroke logging and screen recording with qualitative data from a pre-task questionnaire. De Almeida finds that post-editing performance does not correlate with previous translation or post-editing experience but that it correlates with the participants' attitudes towards machine translation.

Temizöz (2013) compares the performances of subject-matter experts and professional translators when post-editing MT output. She finds "no significant difference between the translators and engineers with regard to postediting and revision speed" but that "engineers produce higher-quality posteditings" (op. cit.: 231). These are very interesting findings with relevant implications for the translation industry when deciding on how to combine the skills of area specialists with those of translators when translating specialised texts. Temizöz presents a comprehensive review of the literature on topics very similar to those covered in the present thesis, namely translation memories and machine translation.[5]

3.3. TM & MT integration

When TM and MT are integrated in the same workflow, new questions emerge: What is the increase in productivity from the resulting integration? How is quality affected? What changes happen in the tasks performed by language professionals?

An early study on the topic is Lange and Bennett (2000), who describe the strategy used by Baan in the late 1990s to integrate machine translation and translation memories

[5] Temizöz (2013) includes two tables summarising the methods used in the reviewed literature: Table 2 (op. cit.: 40), for studies on translation memories, and Table 3 (op. cit.: 66-7), for studies on machine translation and post-editing. The tables include the number and profile of participants, type and length of texts, language direction(s) and the tool(s) used. Although the tables do not include the data collection methods and main results, which are mentioned in the text, they are a very useful source of reference for future researchers.

for the translation of software documentation from English into German. Using a workflow that combines pre-editing, TM matching, rule-based machine translation, macros and regular expressions, they report a reduction of up to 50 percent in translation times, “but only if everything ran smoothly”. Among the conditions for maximising the gains in translation speed are not only the quality of the source text and the MT engine, but also the proper training and motivation of translators (op. cit.: 216).

Another study on the integration of MT and TM is presented by Bruckner and Plitt (2001). The purpose in this case is to find which level of TM fuzzy matching best corresponds to the output produced by an MT system. Although the answer seems to depend on factors such as the characteristics of the texts and the MT engine, the study investigates a crucial question in the translation industry up to these days, i.e. how to pay post-editors for their work on MT suggestions as compared to the established payment schemes for TM matches.

O’Brien (2006a) presents a more robust study on the same topic by introducing eye tracking as a research tool to investigate the translation process. She combines eye-tracking data with screen recordings and think-aloud protocols to compare speed and cognitive effort (measured indirectly through pupil dilation) between exact matches, fuzzy matches, no matches and machine translation. The four translators in her experiment worked in a TM system (Trados Workbench) with translation suggestions coming from Symantec’s legacy translation memories and from a rule-based MT engine (Systran). Although the small data set does not allow firm conclusions to be extracted on the various types of TM matches, the findings indicate a direct correlation between translation times and cognitive effort, with exact matches ranking lowest, no matches ranking highest, and fuzzy matches and MT ranking at intermediary levels of time and effort. O’Brien also concludes that “Machine Translation matches appear to lie in the same region as 80-90% Fuzzy Matches, in terms of cognitive load” (op. cit.: 199–200).

It is worth noting that the three studies mentioned so far in this section carry out their comparisons of TM vs. MT performance within traditional TM systems. This is not the case with other studies, which use post-editing tools for their comparisons. An example is the report presented by Autodesk (2011), where TM and MT are compared in a post-editing environment. Although the methodology is not very clearly explained, the report suggests that the texts translated based on pre-inserted TM suggestions contained fuzzy matches of different levels “including below 50%”, which were presented without the corresponding metadata. Unsurprisingly, the findings of the study indicate that MT post-editing outperforms TM repairing on every account.

Another study that compares TM and MT in a post-editing environment is Guerberof Arenas (2009), who finds that "translators have higher productivity and quality when using machine translated output than when processing fuzzy matches [at any percentage level] from translation memories" (op. cit.: 11). This study does not look into cognitive effort, but in the case of speed its findings contradict those obtained by O'Brien (2006a). Of course, the studies are not directly comparable, as they used different texts, language pairs, MT engines and participants. However, the difference is still considerable enough to suggest that the types of tools used in the experiments (a TM system in one case vs. a post-editing tool in the other case) might play an important role in the results.

As suggested by Figure 9 in O'Brien (2006a), her study used a TM system with a normal configuration for TM matches and a -15% penalty for MT suggestions. This means that translators could know for every segment what type of suggestion they were editing. On the other hand, the post-editing environment used in Guerberof Arenas (2009) presented the translation suggestions with no further information on their provenance (TM or MT) or fuzzy-match level. Based on these observations, I propose that *translation metadata* are an important element that distinguishes TM tools from post-editing tools and are therefore an important variable to be taken into consideration in studies that compare TM and MT. In fact, in a follow-up study, Guerberof Arenas (2012) finds machine translation and fuzzy matches in the 85-94% range to produce similar levels of productivity, and she acknowledges that fuzzy matches could be even faster if the changes in the fuzzy-match segments were highlighted (i.e. if translation metadata were presented for TM suggestions) (op. cit.: 241-2).

Despite the potential relevance of the topic, empirical research on the actual usefulness of translation metadata is still very scarce. Morado Vázquez (2012) approaches the topic by looking at how the presence of metadata affects translators' performance and perception when dealing with TM matches (her study does not include MT suggestions). She compares the behaviour of 33 professional translators working in three different scenarios: without any translation memories (A), with a translation memory but without metadata (B), and with a translation memory and basic metadata elements, mostly project-specific (C). She uses screen recording and keystroke logging to measure translation speed and the LISA QA model to assess translation quality. Her process data indicate that in terms of both speed and quality there is no significant difference between scenarios B and C. In scenario A translators were slower and produced translations of lower quality than in B and C. When it comes to the self-reporting data from the questionnaires, however, most participants indicate that they prefer to have access to the metadata and even believe they can translate faster and better when proper metadata is available (contradicting the actual performance data).

My own pilot experiment used similar data collection methods and compared the performance of professional translators between two environments that contain TM and MT suggestions: one environment presents a selected set of metadata elements and the other presents no metadata. The results indicate a difference in speed and typing activity depending on the types of translation suggestions and the presence or absence of metadata. However, the results varied between the two participants in the experiment, indicating that there might be no single answer as to whether or how particular metadata elements affect a given translator (see Teixeira 2011 and also section 4.4 below).

Morado Vázquez and Torres del Rey (2011) go into the details of which pieces of metadata are most relevant for the translator. Their initial experiment indicates that "some metadata elements are more often taken into consideration [by translators] than others" and that "the usability of the metadata provided has a lot to do with the *way* the tool presents it or works on it" (op. cit.: non-paginated, emphasis added).

Even if only tangentially, Karamanis et al. (2011) find that translation metadata elements such as author, date and revision status affect the way translators attribute *trust* to translation suggestions (op. cit.: 41).

Having explained why translation metadata may need to be taken into account when analysing how translators handle translation suggestions coming from TM and MT, I will now expand the topic to include a larger set of characteristics present in translation tools in general.

3.4. Ergonomics and usability

Ergonomics and usability are aspects of the human-technology interaction that also deserve more attention. Drawing on a survey of 874 translation professionals from 54 countries, Lagoudaki (2006) suggests that "many of the existing commercial TM systems are technology-driven applications (e.g. with an abundance of useless features and a complex, impractical and difficult to learn user interface), rather than user-driven applications" (op. cit.: 4). Although several years have passed since her survey, this seems to still be the case.

However, in recent years, some aspects of the translator's work environment have begun to attract the attention of researchers. O'Brien (2009b) posits that TM systems that present redundant information on the screen might hinder cognitive processing and slow down translation speeds. She offers some suggestions to make the user interfaces more "user-aware", such as "engag[ing] translators in UI design" and designing interfaces with a view to improving usability and "eas[ing] cognitive processing" (op. cit.: 29). A few years later, O'Brien (2012) reinforced those ideas by suggesting that cognitive ergonomics play a more prominent role in the development of translation tools. Other authors have also started to focus on ergonomics as a means of preventing occupational diseases, such as those caused by long exposure to different types of physical strain in the translator's computerised workplace (Ehrensberger-Dow and Massey 2014).

In order to address those issues, TS has started to expand its research scope to include other fields. We need to have an understanding of at least two broad areas: 1) how the human brain processes information when translating, and 2) how the interaction with computers affects human behaviour.

The first point has been brought up in studies such as Christensen (2011), who provides an overview of studies that deal with "mental processes" in the interaction between translators and TM tools. Although some earlier publications did also cover the cognitive aspects of translation (Krings 1986; Danks et al. 1997), this is an area that has received increased attention over the last years, with entire volumes dedicated to the topic (Göpferich et al. 2008; Göpferich and Jakobsen et al. 2009; Shreve and Angelone 2010; O'Brien 2011; Schwieter and Ferreira 2014; Muñoz Martín 2014; Ferreira and Schwieter forthcoming).

The second point has been mentioned by Christensen and Schjoldager (2010), who call for more research "on how translators interact with TM technology and on how it influences translators' cognitive processes" (op. cit.: 99). It has also been addressed by O'Brien (2012), who indicates the need to expand research on translator-computer interaction drawing on the knowledge produced in the broader field of Human-Computer Interaction (HCI), to help understand issues such as how the information on screen (of which translation metadata is part and parcel) affects cognitive processes during translation.

An example of a recent study in the field of HCI with a focus on translation is Green et al. (2013). The authors compare two translation conditions: from scratch (which they call “unaided”) and post-editing (Google Translate). The paper takes into account many factors in addition to the main independent variable (translation condition), including source-text length, syntactic complexity and parts of speech, as well as subject skills and hourly rates. This is made possible by the sophisticated statistical analysis used in the paper (involving linear and ordinal mixed-effects models), which is preceded by thorough consideration of alternative statistical tests. Green et al. (2013) find that post-editing is quicker, produces higher-quality final translations and is a more passive activity than translating from scratch, in the sense that it generates less event counts, longer pauses, and less edits. The results are unsurprising, but the methods of data analysis used in the paper are worth considering. A shortcoming in the study is that it uses mouse hover patterns as an indication of attention allocation and cognitive processing. While the authors refer to other research papers in the field of HCI that corroborate this mouse-attention relationship during web browsing, the same might not apply to translation, since translators typically read in a more discontinuous way (cf. Jakobsen and Jensen 2008) and have to type much more than when just web browsing.

An additional contribution from this paper for the TS community is the method used for assessing translation quality. The researchers used Amazon’s Mechanical Turk crowd-sourcing platform to hire many people to rank pairs of translations, then used a sophisticated formula to re-rank all the translations as a whole. Even though this method originated in the NLP community for ranking MT engines, it has the potential of contributing to human translation assessment in TS studies. Such studies reflect the increasingly necessary interdisciplinarity of research on how translators interact with translation technology.

Having briefly seen the relation between translators and translation technologies from different technical perspectives, let us now consider this relation from a more human perspective.

3.5. The human factor

Previous research indicates that *attitudes* to technology are as important as technology itself (McBride 2009; Morado Vázquez 2012; Doherty and Moorkens 2013). Therefore, increased translation flows are likely to have negative effects if they turn translators into "language soldiers". Several studies have looked into how translators' attitudes affect the adoption and use of translation technologies, such as the surveys conducted by Dillon and Fraser (2006) and Lagoudaki (2006), the questionnaires and debriefings used by Guerberof Arenas (2013) or my own interviews (cf. sections 5.2 and 5.3 and Teixeira 2014a).

Huang (2011) also shows that language professionals do not have the same level of expectation when dealing with human translation as opposed to machine translation. Huang's surveys in the realm of literary translation indicate that 73 percent of the language professionals consider that machine translation can be helpful to translators in some way, either by assisting "a translator in choosing words and sentences to speed up translation" or by "[translating] drafts, leaving editing and proofreading for human translators" (op. cit.: 5, Figure 7). This is in accordance with the results of another survey in the same study that indicates that the role of machines should be to "assist the human translator", "improve humans' efficiency" and "reduce the pain of translating" (op. cit.: 6, Figure 9).

Nyberg et al. (2003: 272) report a similar finding among translators who are required to translate texts written in a controlled language, noting that "[t]ranslators [...] tend to think of their work in holistic terms, and prefer to produce texts which flow from beginning to end with appropriate stylistic variation" (cited in Wallis 2006: 48).

Doherty and Moorkens (2013) investigate attitudes towards TM and MT among students in the context of translation technology teaching. They find that attitudes towards TM tend to be positive, while attitudes towards MT tend to be negative. Looking at these results across time, and comparing with similar surveys on TM a decade earlier, we can assume that students perceive TM as a consolidated technology while MT is still considered as something new, and even potentially threatening.

As has been mentioned in the conclusion to the study by Lange and Bennett (2000), the attitudes of translators towards the technologies they employ can have an impact on their performance when using the technologies. However, in addition to performance, job satisfaction is another important aspect that is affected by translators' attitudes and perceptions towards technologies, as reported by Wallis (2006) and others, and as I will indicate when analysing the feedback from the interviews later in this thesis. For example, as mentioned above, Wallis (2006) found that:

> [...] while productivity seems comparable across the two methods [pre-translation mode vs. interactive mode], the quality of the texts appears to be slightly higher when using interactive translation, and the *job satisfaction of translators is considerably higher when using interactive translation*. (Wallis 2006: 91, emphasis added)

Job satisfaction in general is a complex subject and has been studied extensively in other fields such as Occupational Psychology (Bowling et al. 2010). Job satisfaction within our discipline has been the object of scant research, and is usually associated with surveys of translators' attitudes, as in the few studies illustrated above. One exception is Liu's (2011) thesis on the job-related happiness of 193 translators in the greater China, which unfortunately did not delve into how happiness interacts with translation technologies. In the present thesis I try to contribute to the knowledge in this area by including an analysis of translators' perceptions in relation to the tasks they perform in my translation experiment.

3.6. Process research methods

The current thesis studies the translation process as it takes place, i.e. it investigates what happens while translators translate (as opposed to other approaches that focus on translation as a product). Crucial for studies that focus on the translation process are the strategies and technologies that allow researchers to collect real-time data from the actual translation process, which today typically takes place on a computer. Here I refer back to O'Hagan 2013, mentioned in the beginning of this chapter, when she says that technology has changed both the ways how translation takes place and the possibilities of studying translation.

Alongside the more traditional data-gathering methods such as think-aloud protocols (TAPs) (Ericsson and Simon 1998; Krings 2001; Jakobsen 2003), other methods such as keystroke logging (Jakobsen 2002; 2006), eye tracking (O'Brien 2006a; 2009a) and screen recording allow researchers to identify where attention is being placed and even to measure cognitive load, e.g. through pupil dilation (O'Brien 2006a; Shreve and Angelone 2010). Several studies also combine those methods for better confirmation of results (Alves 2003; Dimitrova 2005; Carl et al. 2011).

The research tools and methods used in this thesis will be discussed further in Chapter 4 (Methodology), but one note might be relevant at this point. Most studies involving translation process research have been done in lab settings. This is due to several reasons, the most prominent of them being the difficulty of observing what translators are doing on the computer in an uncontrolled situation. This thesis, on the other hand, will report on a *workplace* experiment, in an attempt to come closer to the real world of translators. This brings in additional challenges, as has been reported in studies such as Séguinot (2000) and Asadi and Séguinot (2005).

Ehrensberger-Dow (2014) is a more recent example of translation process research at the workplace that employs methods similar to the ones I use in this thesis: "interviews, questionnaires, computer logging, […] screen recordings […], eye-tracking and retrospective verbalizations" (op. cit.: 360) as well as "translation evaluation" (op. cit.: 366) and triangulation between qualitative and quantitative data. Ehrensberger-Dow mentions issues similar to those considered in my own work, such as how to recruit participants, how to choose source texts, confidentiality and ethical aspects. Her experiment was on a larger scale, with more participants, more researchers and spanning over a longer period. It was also broader in scope, as it looked not only at the translation act as it unfolds, but it also included other phenomena that take place over the life of a translation project, such as the interaction between translators and project managers.

Ehrensberger-Dow (2014) illustrates the many challenges associated with workplace studies. For example, severe confidentiality issues were at stake, to the point that keystroke logging could not be used at all (op. cit.: 371-2) and had to be compensated for with the available screen recordings. Ehrensberger-Dow also faced difficulties with the use of eye tracking. In fact, she could not use the resulting data either, because of security regulations related to one eye-tracker model and low data accuracy with another model (op. cit.: 375-6). She suggests that "[n]ewer models of eye-trackers, such as those that can be installed under an existing monitor" be used as an alternative. Coincidentally, this is exactly the type of eye tracker I used in my workplace study, which also presented many challenges, as will be discussed later in this thesis.

3.7. Conclusion

As one would expect, empirical studies with a focus on translation memories (Webb 1998; Dragsted 2004; Colominas 2008; Moorkens 2012) have reported on the use of typical translation memory systems. These are tools that offer one or more translation suggestions as the user activates a segment and that always display metadata about those suggestions, i.e. they indicate where the suggested translations come from, how similar to the reference source segment the current source segment is (fuzzy match level) and where the textual differences lie. In contrast, studies on pure machine translation post-editing (Krings 2001; Guerra Martínez 2003; Allen 2003; Garcia 2010; Plitt and Masselot 2010; de Almeida 2013) have often resorted to editing environments that offer pre-translated text with no associated metadata, as this is the typical setup for such tools.[6]

Some studies have compared unaided human translation with TM-assisted translation or with MT-assisted translation. However, only a few studies have analysed scenarios in which machine translation and translation memories are combined in the same workflow. These studies either use existing TM systems (O'Brien 2006a; Skadiņš et al. 2011; Yamada 2011) or they resort to a purpose-built post-editing environment (Guerberof Arenas 2009; He et al. 2010), as there seem to be no established tools for post-editing. One question that arises from this dichotomy is how to compare the performance of TM suggestions against MT suggestions in an environment that has not been conceived with their integration in mind. On the one hand, in a post-editing tool TM matches are analysed without the associated metadata, which are an important feature of translation memory systems (Morado Vázquez and Torres del Rey 2011; Karamanis et al. 2011; Morado Vázquez 2012; Teixeira 2014b) but are not present in post-editing tools. Metadata could help translators not only to make choices among different types of suggestions, but also to decide how to approach a suggestion when repairing it. On the other hand, in a traditional TM system, MT suggestions have to be manually inserted in the active segment and are presented surrounded by much more information than is typical in a post-editing tool, maybe decreasing the translation speed for this suggestion type and increasing the post-editor's cognitive load. Therefore, comparing the performances of TM vs. MT suggestions is not an easy task, as the general tendency is to assess one of the suggestion types in an environment for which it was not originally intended to be used. For these reasons, none of the previous research convincingly

[6] The scenario for post-editing is starting to change with the development of post-editing environments that can display confidence estimates for machine translation suggestions, such as PET (Aziz et al. 2012) and CASMACAT (2014). Those estimates are believed to represent useful metadata for repairing MT suggestions.

answers this simple question: Does the presence of metadata help or hinder translation processes that work with TM and MT?

Since there is very little research on this topic, this thesis seeks to validate the assumption that translation metadata is an important element that distinguishes translation memory workflows from post-editing and revision workflows. The study reported on here uses a traditional TM system, but the system is set up using different configurations, in an attempt to “favour” one suggestion type at a time. In addition to the quantitative effects of metadata on translators’ performance, I will also analyse the translators’ perceptions and preferences related to the different tasks proposed, focusing on the varying configurations of the tool interface.

Chapter 4. Methodology

In this chapter, after presenting my research question, hypotheses and variables, I describe a pilot experiment I ran before the main experiment. Much was changed between the pilot and the main experiments, due to the reasons that will be explained in section 4.5.1. The pilot study used a mainstream TM system (SDL Trados Studio 2009) and a generic MT engine (Google Translate), while the main study used a less widespread TM system (IBM TranslationManager) and a customised MT engine (based on Moses). In addition, different data collection tools were used, including an eye tracker in the main experiment. Those differences between the two studies explain the great level of detail used to describe the pilot study in this thesis.

After presenting the pilot experiment, I describe the materials and procedures of the main experiment, and then I explain in more detail the data collection methods and the equipment involved. Later in the chapter I explain how I analysed the data, focusing on each of the dependent variables.

In both the pilot and my main study, my empirical approach was experimental. It could not have been simply observational, since in order to gather data with the level of detail required to answer my question I needed to resort to research tools that could not be installed without the participants' knowledge or consent. In both cases, the experimental setting was as close as possible to the translators' normal work environment, in an attempt to increase ecological validity as much as possible.

Most of the data gathered from the experiments are of a quantitative nature, although I try to relate those with the qualitative data gathered from the interviews. For this purpose I will resort to mixed-methods approaches (Johnson et al. 2007) for some of the analyses.

4.1. Research question

My main research question can be summarised as follows: "What is the impact of translation metadata on translators' performances?". It could also be rephrased as: "What are the differences (if any) in the translation process between a situation where translators know the metadata about the translation suggestions and a situation where the metadata are not available?".

One set of reasons for asking this question revolves around understanding how translators behave under both conditions, in order to come up with best practices for working in scenarios that combine translation memories (TM) and machine translation (MT), and to investigate some of the cognitive and emotional factors involved. Another set of reasons stems from the need to investigate a *method* for comparing the usefulness of both suggestions (TM and MT) under similar conditions, with a view to assessing to what extent the results of studies that analyse translation-with-metadata are comparable to those of studies that report on translation tasks without metadata, as explained in section 3.7 above.

4.2. Hypotheses

In order to answer my research question, I compared two translation tasks, in both the pilot and main experiments. In one of those tasks, the translators could see the metadata on the translation suggestions (Visual task), whereas in the other task they did not have access to this information (Blind task). These are my working hypotheses:

- Hypothesis 1 (H1): The presence of *metadata* affects *translation time.*
  - Sub-hypothesis 1a (H1a): The effect of *metadata* on *translation time* varies in accordance with the *type of translation suggestion.*
- Hypothesis 2 (H2): The presence of *metadata* affects *typing effort*.
  - Sub-hypothesis 2a (H2a): The effect of *metadata* on *typing effort* varies in accordance with the *type of translation suggestion.*
- Hypothesis 3 (H3): The presence of *metadata* affects *error scores*.
  - Sub-hypothesis 3a (H3a): The effect of *metadata* on *error score* varies in accordance with the *type of translation suggestion.*

Some definitions are necessary in order to operationalise the variables I want to test.

4.3. Definitions of variables

4.3.1. Translation time

Translation time is indicated as seconds per 100 words of source text. It is measured using keystroke logging tools from the moment the translator activates a segment to the moment the translator closes the segment.

4.3.2. Typing effort

Typing effort is indicated as the percent ratio between the number of keystrokes performed by the translator while editing a particular segment and the total number of characters in the resulting segment. It is also measured using keystroke-logging tools. For example, if a translator types 35 characters to produce a 100-character long translation (because some characters will have been leveraged from the translation suggestion), the typing effort for that segment is 35 percent. Deletions are also counted as keystrokes (see section 4.8.1 for details).

4.3.3. Error score

Error score is indicated as errors per 100 words of source text. It is obtained from a quality assessment done by human reviewers (see section 4.6.5).

It is worth noting that the above three variables are all relative to text segment length. This was done to allow direct comparisons between segments of different lengths when analysing each variable. Time and errors are calculated per 100 source words, while typing effort is normalised on the basis of the target segment length.

The decision of choosing the source text as the reference for time and errors was based on the standard practice in the industry. Not less important, the choice was based on my intuition that the origin of the problem-solving strategies lies within the source text, even recognising that the translation suggestions work like a second source text and turn decision-making into an even more complex process.

On the other hand, I used the target text for measuring typing effort in order to account for the differences in length between the source and target languages. For example, if a source segment has 30 characters and the final translation has 34 characters, which were produced by typing only two characters on top of the initial translation suggestion, the variable as I calculate it has a value of 2/34 = 5.88 percent. If instead we considered the source segment as the reference for calculating the typing effort, the result would be 2/30 = 6.67 percent. I believe the first method reflects better the work that has actually been done by the translator.

It should also be noted that the variables were defined so that an increase in their values indicates phenomena going in the same direction, i.e. “the higher the worse”. This will make it easier to interpret the data in graphs and tables.

4.3.4. Translation metadata

"Translation metadata" is defined as the combination of the following elements displayed in the translation tool: type of origin, translation-memory name, last usage date, match type, fuzzy match level (%) and textual differences. This is a binary variable: metadata are either present or absent. See section 1.4 and Appendix 11 for an explanation of the concept of translation metadata.

4.3.5. Type of translation suggestion

For each segment in the translation tasks, translators receive a translation suggestion, which can be of four different types:

- a translation-memory exact match (Exact Match);
- a translation-memory fuzzy match in the 85-99% range (High Fuzzy Match);
- a translation-memory fuzzy match in the 70-84% range (Low Fuzzy Match);
- a machine-translation feed (Machine Translation).

The word "suggestion" has been chosen instead of alternatives such as "proposal", after consultation with translation scholars that are native speakers of English and on Internet forums. The general consensus is that "proposal" involves some commitment and tends to be more associated with human action. Since I wanted a term for both TM and MT-generated suggestions, this was the preferred term.

4.4. Pilot study

Prior to the main experiment, a pilot experiment was carried out. It tried to answer similar research questions by comparing a task with translation metadata (at that time called "provenance information") with a task without translation metadata.

4.4.1. Research question and hypotheses

I set out to investigate whether the fact of knowing the "provenance" of the segments could provide an explanation for apparently contradictory findings related to productivity in some studies. My initial research question was: "What are the differences (if any) in the translation process between a situation where translators know the provenance of the translation suggestions they are editing and a situation where this information is not available?".

In order to answer that question, I compared two translation environments. In the first environment, translators did not know the provenance of translation suggestions, whereas in the second environment translators did have access to this information. The same translators completed both tasks in alternating orders.

These were my working hypotheses:

- Hypothesis 1 (H1P): The *translation speed* is higher when *provenance information* is available.
- Hypothesis 2 (H2P): There is no significant difference in the *quality level* when *provenance information* is available.

"Translation speed" was initially measured as words per hour. To make it easier to compare this variable with the corresponding variable in the main experiment, the results will be presented in the same unit used for "translation time", i.e. in seconds per 100 words. The "quality level" was measured as a score given by two reviewers, who processed all resulting translations according to predefined criteria. The "provenance information" of translation suggestions was indicated by showing their origin (TM or MT) and, in the case of TM, by displaying their fuzzy-match percentage and highlighting the differences between the actual segment and the matching segment in the TM. This corresponds to what I now call "translation metadata".

4.4.2. Participants

There were only two participants, both men and both L1 speakers of Spanish. P01p had had formal training in translation and four years of professional experience in several fields, especially in audio-visual translation. P02p had also had formal training in translation and around eight years of professional experience in various fields, mainly in localisation and technical translation. Both were doctoral students of Translation Studies and were familiar with many different translation memory systems.

The reviewers were teachers of translation in the Department of English and German Studies at Rovira i Virgili University.

4.4.3. Experimental setup

The two translation environments were created within SDL Trados Studio 2009 Freelance. The source texts were taken from an article in a technical magazine and dealt with composite materials in car manufacturing. The specific text was chosen based on its type (technical marketing on a topic of general interest) and length – allowing for the extraction of two excerpts of around 500 words. The main article had a total of 1310 words, corresponding to 55 source segments, or 23.8 words per segment in average. In order to have two source texts of around 500 words, I used 21 segments for each of them. As a result, one source text had 512 words, and the other had 510 words. A translation memory was created by aligning the English source text with the Spanish target text (the final version approved by the client) using SDL Trados WinAlign, plus manual verification of each segment. The aligned translation memory was edited to produce two smaller memories (one for each text) with the following distribution of translation suggestions:

- 7 “no matches” (replaced with MT feeds);
- 5 exact matches;
- 9 fuzzy matches, of which:
    - 3 matches within the 70%-79% range,
    - 3 matches within the 80%-89% range, and
    - 3 matches within the 90%-99% range.

The order of presentation of match types during the translation was defined by a random-number generator and it was different for each of the tasks. Segments set to have an “exact match” suggestion were left untouched. Segments corresponding to a “no match” were replaced through SDL Trados Studio with translation suggestions provided by the public, (at that time) freely available Google Translate machine-translation service. Finally, for creating the fuzzy matches I resorted to the following strategies: delete parts of the source and target segments in the TM, include or replace some words in the source and target TM segments, or edit the source text.

As in the main experiment, ecological validity was a major concern in the pilot experiment. Both participants used their own laptop computers and worked in a small classroom at Rovira i Virgili University on different days. The aim was to have the translators work in a setting as close as possible to the natural work environment of a freelance translator, meaning that they could keep their preferred configuration in terms of keyboard, screen and mouse (either built-in or external), operating system (within the Microsoft Windows family), browser favourites, dictionaries, etc. They also had access to the Internet during the experiment. Before they started, we made sure they had the required versions of SDL Trados and BB FlashBack installed and configured. The participants were briefed on the main goals of the experiment and they signed release forms giving their informed consent.

The translators were given instructions in Spanish on how to perform the main tasks for the experiment. In general terms, the instructions indicated that the translation memory had been created based on a client-approved final version of the Spanish magazine, that it contained five different kinds of matches, and that machine translation was being used to replace "no match" segments. The translation instructions also mentioned that the translators should act as if they were going to be paid the same amount per word (no fuzzy-match discounts), thus implying that they were supposed to revise all segments, including exact matches. No actual payment was offered; the participants were volunteers. The instructions made it clear that their translations were going to be assessed and graded for quality by a professional reviewer, implying that the translators should try to achieve maximum quality in both environments. A time limit of 1.5 hours was set for each of the texts.

4.4.4. Data collection

The main methods for collecting data were screen recording and keystroke logging with BB FlashBack Express 2. Retrospective interviews were also used to try to obtain some insight into the translators' feelings and satisfaction in both tasks. For testing quality, all texts were rated by two reviewers: first based on an error-count system using a score that started at 10 and decreased according to a predefined grid, and then holistically, giving a score for the overall quality of the translation as a text.

At the beginning of the experiment, a digital voice recorder was turned on. During the translation of the texts in both environments, BB FlashBack was set to record the following data: screen activity; keystrokes; mouse position, movements and clicks; translators' faces; and sound (voices, keyboard, etc.).

Time was measured by watching each of the translators' performances in BB FlashBack Player and by manually noting down the start and end times for each individual segment. Time was counted when translators were typing, thinking, hesitating or looking at the source text (except when they read the full source text before starting the translation, as it would not be possible to make a correspondence between that time and specific segments). Time was not counted when translators switched to another window to look up terminology, tried to find a specific function in the tool or spoke with the researcher, since those activities were not the object of the experiment. The time spent on searches within the translation environment (mainly with the Concordance function) was considered as translation time.

4.4.5. Results

During the experiment, although there were no specific instructions in this regard, both participants translated the entire text sequentially, each segment at a time (the *drafting* phase), then they read through the entire text again (the self-*revising* phase). After encountering this phenomenon, I decided to organise and present the data separately for the two phases.

4.4.5.1. Time

Table 1 and Table 2 show the time results for Participant P01-P in the two tasks.

Table 1. Average relative times per type of suggestion for P01-P in the Visual task (seconds / 100 words)

Type of suggestion	Drafting	Revising	Combined
Exact (100%) matches	119	71	190
90-99% matches	258	111	369
80-89% matches	301	52	353
70-79% matches	461	101	562
Machine translation	522	88	610
Whole text	339	86	425

Table 2. Average relative times per type of suggestion for P01-P in the Blind task (seconds / 100 words)

Type of suggestion	Drafting	Revising	Combined
Exact (100%) matches	442	39	480
90-99% matches	567	82	649
80-89% matches	273	39	312
70-79% matches	355	26	381
Machine translation	359	78	436
Whole text	392	55	447

If we look at the results for the first phase (drafting), we see that the mean translation time for the whole text is lower in the Visual task (339 seconds / 100 words) than in the Blind task (392 seconds / 100 words), a difference of 15.6 percent. If we look at the results for the first and second phases (drafting + revising) combined, the translation times are still slightly lower in the Visual task (425 seconds / 100 words) than in the Blind task (447 seconds / 100 words), but the difference is reduced to 5.2 percent. This indicates that the Blind task required proportionally less time for revising than the Visual task did.

If we move away from the entire text and look into the five groups of segments, with their different types of translation suggestions, it is possible to identify internal differences in time. This is in accordance with intuitive expectation and with the results obtained by O'Brien (2006a).

Moreover, when the two tasks are compared, there is a dramatic increase in time for exact matches (from 190 to 480 seconds / 100 words) in the Blind task, suggesting that translation metadata have a high impact for this kind of translation suggestions. Matches in the 90-99% range also show a dramatic increase in time (from 369 to 649 seconds / 100 words), again indicating that metadata have a significant impact in this case. Matches in the 80-89% range did not show a relevant variation. For lower fuzzy matches and MT feeds, it is worth noting that there was a *decrease* in time. This means that the translator spent less time translating low fuzzy matches and MT feeds when he did not know the type of suggestion he was editing.

Table 3 and Table 4 show the average time results for Participant P02-P in the two tasks.

Table 3. Average relative times per type of suggestion for P02-P in the Visual task (seconds / 100 words)

Type of suggestion	Drafting	Revising	Combined
Exact (100%) matches	180	92	272
90-99% matches	389	176	565
80-89% matches	442	152	594
70-79% matches	524	156	680
Machine translation	316	143	459
Whole text	342	139	481

Table 4. Average relative times per type of suggestion for P02-P in the Blind task (seconds / 100 words)

Type of suggestion	Drafting	Revising	Combined
Exact (100%) matches	348	74	422
90-99% matches	422	80	503
80-89% matches	293	186	478
70-79% matches	375	102	477
Machine translation	344	98	442
Whole text	352	104	455

For this translator, the results for the first phase (drafting) show that the mean translation times were also shorter in the Visual task (342 seconds / 100 words) than in the Blind task (352 seconds / 100 words), but the difference is much smaller than for participant P01-P, at only 2.9 percent. The combined results for the first and second phases (drafting + self-revising) show that translation times are now shorter in the Blind task (455 seconds / 100 words) than in the Visual task (481 seconds / 100 words), with a difference of 5.7 percent. Although there is insufficient data for meaningful statistics, this difference does not appear significant.

Now let us look again at the time differences according to the various suggestion types. Roughly speaking, the data for the Visual task indicate that P02-P processed translation suggestions coming from exact matches at half of the time spent for suggestions coming from fuzzy matches, and he spent less time to translate suggestions coming from machine translation than from fuzzy matches. The shorter times for exact matches are in accordance with my expectations, but the reasons for machine-translation suggestions being translated in less time than high-percentage fuzzy matches should be investigated further.

In the Blind task, similarly to what happened with P01-P, the data for P02-P indicate a dramatic increase in the average translation times for suggestions coming from TM exact matches (55.2 percent, from 272 to 422 seconds / 100 words). All other kinds of translation suggestions had a decrease in time. It is interesting to note that differences in translation times tend to disappear in the Blind environment: exact matches were translated at slightly shorter times, at 422 seconds / 100 words, followed by machine-translation suggestions, at 442 seconds / 100 words, with translation-memory fuzzy matches taking a little longer, between 477 and 503 seconds / 100 words. The differences between the five types of translation suggestions do not appear to be significant.

4.4.5.2. Quality

Two revisers assessed the quality of the four translations (two per subject) using a predefined grid. The revisers were then told to compare the two translations from the same subject and decide which one was better, if any, and to give their final grade from 0 (worst) to 10 (best). This means each reviser scored the translations twice – once according to the grid, then again holistically. The results are shown in Table 5. We should keep in mind that the method used for quality assessment in the pilot experiment was different from the method used in the main experiment, and that the numbers in Table 5 represent a “lack of errors”, instead of an error score (so, here, the higher the number, the better).

Table 5. Translation quality levels for both participants in the pilot experiment

	P01-P		*P02-P*	
	Text 1 (Visual)	Text 2 (Blind)	Text 1 (Visual)	Text 2 (Blind)
Reviser 1	8.5	7.0	8.5	9.0
Reviser 2	7.5	7.0	8.0	8.5
Average	8.0	7.0	8.25	8.75

According to the two evaluators, P01-P performed better in the Visual task, while P02-P performed slightly better in the Blind task.

4.4.6. Discussion

The results from the pilot experiment did not allow me to draw a definite conclusion on my first hypothesis (on translation times). P01p had slightly shorter times (4.9 percent) in the Visual task, while P02p had slightly shorter times (5.4 percent) in the Blind task. However, the overall speed, besides individual-specific differences, depends on the distribution of different types of translation suggestions in the texts, as the results showed that translators spent much longer repairing exact matches when they did not have the translation metadata.

As for my second hypotheses, although it was not rejected, it was not possible to determine with a reasonable level of certainty whether the translations produced in the two tasks could be considered of same quality for each individual participant. Moreover, the quality assessment was only done on the texts as a whole, so it was not possible to associate quality levels with specific types of translation suggestions.

Although inconclusive, the results of the pilot experiment still indicated that translation metadata correlate with certain changes in performance. More important, the pilot study allowed me to identify several limitations in the research design and pointed to some changes to be implemented in the main experiment. Here is a list of shortcomings in the pilot experiment and some solutions that were identified:

Few and irregular segments. The larger text from which the source texts were extracted had some very long segments, which obliged me to use only a few segments per type of suggestion in order to avoid increasing the total word count. There was also a high degree of variation in the length of segments, with the shortest segment being six words long and the longest being 44 words long. This made it hard to establish comparisons between the segments, among other reasons because MT is known to perform worse with segments containing extremely short or long sentences (Plitt and

Masselot 2010: 12). For the main experiment, I decided to be more careful when selecting the source texts, to have less variation in the segment lengths (cf. section 7.4.5).

Terminology. Even though the time used for terminology search was discounted, the time spent within the translation tool was higher when the terms were more complicated. This was partly compensated for by the fact that the type of suggestion for each segment was defined randomly, but in order to eliminate extraneous variations, problematic terms should be avoided or a glossary should be provided for them.

Segment identification. When trying to calculate how much time the translators had spent in a specific segment, sometimes it was difficult to identify which segment the translators were focusing on at a given moment. This was especially the case in the self-revising phase, as the segment one is working on does not necessarily correspond to where the mouse pointer or the cursor is on the screen. Eye tracking would be considered as an additional data-collection method to help solve this issue.

Quality assessment. From the evaluators' feedback, I considered that the quality assessment was not done properly and their grades did not make it possible to draw any firm conclusions. One obvious lesson was that the rating instructions needed to be made clearer. In order to establish correlations between productivity and quality, it would be necessary to have a quality indicator for each segment, not only for the full texts. A method would need to be created to allow for a per-segment quality assessment.

Validity of hypotheses. It became evident that my initial hypotheses were too general, as they (implicitly) concerned the full texts. The conclusions and the results of testing the hypotheses should depend on the distribution of suggestion types in the text. It was decided that the hypotheses would have to be subdivided according to suggestion types.

Manual methods. The method used for calculating the time spent in each segment, by manually noting down the start times and end times from BBF recordings, was very time consuming and error-prone. I decided to find a more automated way of recording times (and keystrokes) for the main experiment, especially considering that it would involve more participants and more segments.

In conclusion, the pilot study corroborated my initial assumption that “metadata” have an impact on certain performance metrics and suggested that this impact should be investigated in more detail, based on the different types of translation suggestions.

4.5. Main experiment

In the main experiment, I tried to implement several improvements from the lessons learned in the pilot study, as mentioned above. In addition, I included "typing effort" as a dependent variable and "type of translation suggestion" as an independent variable. At the same time, other changes needed to be made in the research design, to account for the conditions that will be explained below. For instance, instead of using a generic machine translation engine, I used a customised one; instead of using technical marketing texts, I extracted my source texts from a software manual; and instead of using Trados as the translation tool, I used IBM TranslationManager.

4.5.1. Background

Finding participants with a comparable profile for running a translation experiment is a recurrent problem in our field. Many studies have resorted to translation students to palliate this difficulty, as hiring professionals is even more complicated. Conveniently enough, the main experiment that served as the basis for analysis in this thesis had the chance of being run as part of a research placement in a translation company that was a partner in the wider European project TIME, of which I was a grant holder. The possibility of running an experiment in that company with their "real translators" was an excellent opportunity, so I set out to investigate all the existing conditions in order to prepare the experiment.

The company in question is called MSS.[7] It is based in Barcelona, with one office and 15 in-house employees, plus a network of freelancers. Given the small size of the company, I was able to interact with staff at all levels, from the owner to the translators. I had interviews with the project managers and some of the translators, after which I realised that only a few of them were familiar with Trados, for example. More translators were familiar with Wordfast, but I had decided to have not less than ten participants, and there were not ten translators that worked with Wordfast either. I also wanted to have an even distribution of men and women, in order to test for any gender differences. Soon it became clear that IBM was their biggest customer and that only the IBM projects would allow me to find enough participants with experience translating the same type of materials, using the same tool and post-editing machine translation. IBM was also the

[7] http://www.mss.es

only client for which the company had full control over the integration of machine translation, through an online tool set up for the company by their machine-translation provider Tauyou.[8] When a new translation package arrives from the client, the project manager in charge sends the untranslated segments through the machine-translation service and then integrates the machine-translated segments into the translation package that is sent out to the translators. With other clients, MSS has no control over the machine-translation process.

At the same time, the company's owner mentioned my research to the local management at the IBM Translation Centre in Barcelona, who in turn mentioned it to the corporate management in the United States. A conference call was held on 10 May 2012 between the different IBM offices, the translation company's management, my thesis supervisor and myself. Research topics were discussed, with some follow-up emails in the following weeks, and authorisations were granted for me to use their materials for my research purposes. All the necessary conditions had thus been created for me to start planning the actual experiment as a whole IBM "package": source text, translation memories, customised machine translation engine and translation memory system.

4.5.2. IBM projects

IBM projects are usually classified according to the type of material to be translated, which in turn depends on the intended use of the material. The first type of project involves the translation of strings for the Graphical User Interface (GUI) and online documentation (Help files). The second type is called "PUB" (as in Publications) and encompasses manuals and user guides. Finally, the third type of project comprises marketing material. The different types of projects involve the translation of different file types, somewhat different workflows and have different quality expectations, with marketing projects being the most demanding in this regard.

The project set up for the experiment was of the PUB type. These projects can be composed of files of several types, which are translated with IBM TranslationManager (see 4.5.3). After finishing the translation, translators are expected to carry out regular quality assurance (QA) procedures, such as running the spell checker in IBM Translation-Manager and performing several QA checks in the external tool Xbench.

[8] http://www.tauyou.com

In order to integrate MT into the translation workflow, the user (usually a project manager or "file handler") sends the relevant segments for pre-translation to the machine-translation service and then uses the resulting translated segments as a regular translation memory in the folder (a translation project, in IBM's jargon). Translation suggestions from segments that are pre-translated through this process will always contain an "m" flag to indicate they come from MT and are not regular TM matches.

4.5.3. The translation tool

IBM TranslationManager was one of the first translation memory systems to be created in the early 1990s and has been used in production for IBM translations since then. It was initially developed for the IBM OS/2 operating system and is still referred to as IBM TM/2. Its graphical user interface has remained virtually unchanged since its first Windows version in the late 1990s.

As far as the display of translation metadata (see section 1.4) is concerned, IBM TM/2 displays the information on the type of origin (provenance metadata) with a letter ("m" for MT and "f" for TM fuzzy matches) and optionally with colour codes. When a translation suggestion comes from a translation memory, the tool displays the following additional metadata elements: the *file name* from which a translation suggestion was produced, the *date* when a translation segment was last used, the *fuzzy match level (%)* and indications of *textual differences* between the source segment being translated and the source segment(s) of the translation memory or memories from which translation suggestions were produced (see Figure 1 and Figure 2 on pages 57-58).

The graphical user interface of IBM TM/2 allows for extensive customisation, so translators can choose to display more or less metadata on screen. I did not predetermine the elements that they should select prior to the experiment. Nevertheless, the translators chose virtually the same elements, either because they used the default tool configuration or because they had previously been told to select specific options when working with regular IBM projects.

4.5.4. Participants

The ten translators who took part in the experiment were selected based on the suggestions made by the company's production and vendor manager and according to their availability during the period defined for the experiment. In order to be eligible to participate, they needed to have at least one year of experience working as a translator, they needed to work with IBM projects on a regular basis (which automatically meant they were familiar with the translation tool) and they needed to have prior experience with MT post-editing. To make sure the participants met the required criteria, they were asked to answer a questionnaire, which was exchanged via email within the company.

The selected participants were native speakers of Spanish, with some of them being bilingual Spanish/Catalan speakers. There were five men and five women, with ages ranging from 24 to 51. They had been working for 1.5 to 18 years as full-time translators for MSS. They had been translating IBM material and using IBM TM/2, the translation memory system used in the experiment, during their employment at the company. They all had experience post-editing machine translated texts for IBM and/or other customers for 0.5 to 3 years. As a compensation for performing the tasks in the experiment, they were paid their regular hourly rates. Table 6 shows the demographics of the experiment participants.

Table 6. Demographic data of participant translators in the main experiment

Participant	Gender	Age	Years working as a translator	Years working with IBM TM/2	Years working with MT post-editing
P01	F	30	7	6	0.5
P02	M	37	14	13	0.5
P03	F	32	3.5	3	0.5
P04	M	26	2.5	2	2.0
P05	F	26	3	3	0.3
P06	M	29	2.5	2	0.5
P07	F	24	1.5	1.5	1.0
P08	M	51	18	18	0.8
P09	F	43	10	10	3.0
P10	M	47	15	14	0.5

4.5.5. Source texts

The source texts used for the three translation tasks were excerpts from the *Troubleshooting Guide* for the IBM Tivoli Monitoring software, which had around 110,000 words in total. One text of 52 words was extracted from the official Spanish translation of the manual for a preliminary Copy task. This text was named *SourceText_Copy*. Another text with 118 words was extracted from the English version of the manual and used for a second preliminary task where the translators had to translate from scratch. This text was named *SourceText10*. For the main translation tasks, two other excerpts of the same manual were chosen, one with 542 words (named *SourceText31*) and another one with 505 words (named *SourceText42*), each with 28 segments.

IBM manuals have a standardised structure and are supposed to follow controlled language guidelines (Bernth 2006). I wanted to make sure the two larger source texts were comparable in terms of complexity and translation difficulty. At the same time, I wanted both source texts to be representative of the larger manual from which they were extracted. After analysing the specific manual I was planning to use, I identified that it could be divided into two different parts, according to their text structures. The first one corresponded to the main body of the manual, which was composed of (in each particular section or subsection):

- an introductory paragraph comprising one or two sentences, followed by
- a set of instructions on how to perform an action.

Example:

Subscribing to IBM support notifications

You can subscribe to e-mail notification about product tips and newly published fixes through the Support portal. […]

Procedure

1. Open the http://ibm.com website and select Support & downloads > Technical support. You can also launch an IBM support website, such as http://www.ibm.com/support/us.
2. In the Quick start page or Support home, click Sign in to sign in or to register if you have not yet registered.
3. […]

The second part of the manual in terms of text structure was a large Glossary section towards the end of the publication, containing around 6,700 words. The typical text for the Glossary was:

- an initial sentence in the form of a noun phrase, followed by
- one or more sentences, with varied syntactic structures and varied lengths.

Example:

> **client/server architecture**
> An architecture in which the client (usually a personal computer or workstation) is the machine requesting data or services and the server is the machine supplying them. Servers can be microcomputers, minicomputers, or mainframes. The client provides the user interface and may perform application processing.

Both source texts were created so that each of them contained 15 segments of the first type (main body) and 13 segments of the second type (glossary). Segments were chosen from similar sections within the manual in order to produce comparable source texts. For SourceText31, the shortest segment contains 7 words, while the longest one contains 41, with an average of 19.4 words per segment. For SourceText42, the shortest segment also contains 7 words, while the longest one contains 34, with an average of 18.0 words per segment. Tags (mark-up codes) within the text were not removed, as in real translations they are always present. Standard quantitative measures of text complexity were used to confirm that both texts were of comparable complexity, as shown in Table 7.

Table 7. Quantitative indicators of text complexity for SourceText31 and SourceText42

	Flesch Kincaid Grade Level	Flesch Reading Ease	Lexile Measure
SourceText31	12.5	39.1	1410L
SourceText42	12.6	37.8	1470L

The texts used for the two preliminary tasks (Copy and Scratch) and for the two main translation tasks (Visual and Blind) can be found in Appendices 1 to 4.

4.5.6. Translation memory

Each of the 28 segments in SourceText31 and in SourceText42 was randomly and evenly assigned one of four possible types of translation suggestions, resulting in seven translation suggestions of each type per text:

- Seven exact matches (E),
- Seven high-percentage (85-99%) fuzzy matches (H),
- Seven low-percentage (70-84%) fuzzy matches (L), and
- Seven machine translation feeds (M).

An authentic IBM translation memory was used as a reference for producing the exact and fuzzy matches. No special tricks were inserted intentionally, nor were any corrections made to existing typos or inconsistencies. The fuzzy matches were produced by editing the translation-memory source or target segments, as indicated in Appendix 5 and Appendix 6.

The machine-translation feeds came from a commercial Moses (Koehn et al. 2007) statistical engine that had been trained with product-specific terminology and was used in production for regular IBM projects in the company. The segments that needed to be machine translated were sent through an online system, which returned a file with the machine translations in the form of a translation memory.

The TM/2 folder (project) that was used in the translation tasks thus contained two translation memories: one with the translations produced from the exact and fuzzy matches and another one with the translations generated through the MT system. For the translators, the process of retrieving suggestions from those memories was transparent and automatic whenever a segment was activated, but the suggestions originated from the MT process always displayed an [m] indicator.

4.5.7. Running the experiment

The experiment took place from the 5th to the 19th of July 2012 in Barcelona, in the premises of the translation company MSS. Prior to the experiment with each translator, I installed the necessary software on their computers and set up the research equipment at their desks, as explained in section 4.7.

4.5.7.1. The translation tasks

The participants were asked to complete a preliminary task, which consisted in typing a given short text (52 words) in Spanish into any text editor of their choice (the *Copy* task). The text used for this task was an excerpt from the translated version of the same manual used for the translation tasks. The participants were told that the goal of the task was just to ensure that all the research equipment was working as expected. While this was partly true, the Copy task was also meant to serve as a warm-up (and cool-down) activity, so that the translators could get used to the experiment conditions, and to measure the translators' baseline typing performance (and eventually assess whether this had an influence on their editing strategies).

After completing the preliminary Copy task, each translator was asked to perform the following three tasks:

1) Translation from *Scratch*: To translate a short text (118 words, 5 segments) from English into Spanish in IBM TranslationManager, without any help from translation memories or machine translation.

2) Translation in a *Visual* setting: To translate a longer text (505-542 words, 28 segments) from English into Spanish in IBM TranslationManager, with one translation suggestion (from TM or MT) per segment and metadata about the translation suggestions.

3) Translation in a *Blind* setting: To translate a longer text (505-542 words, 28 segments) from English into Spanish in IBM TranslationManager, with pre-translated segments (from TM or MT) but no metadata about the translation suggestions.

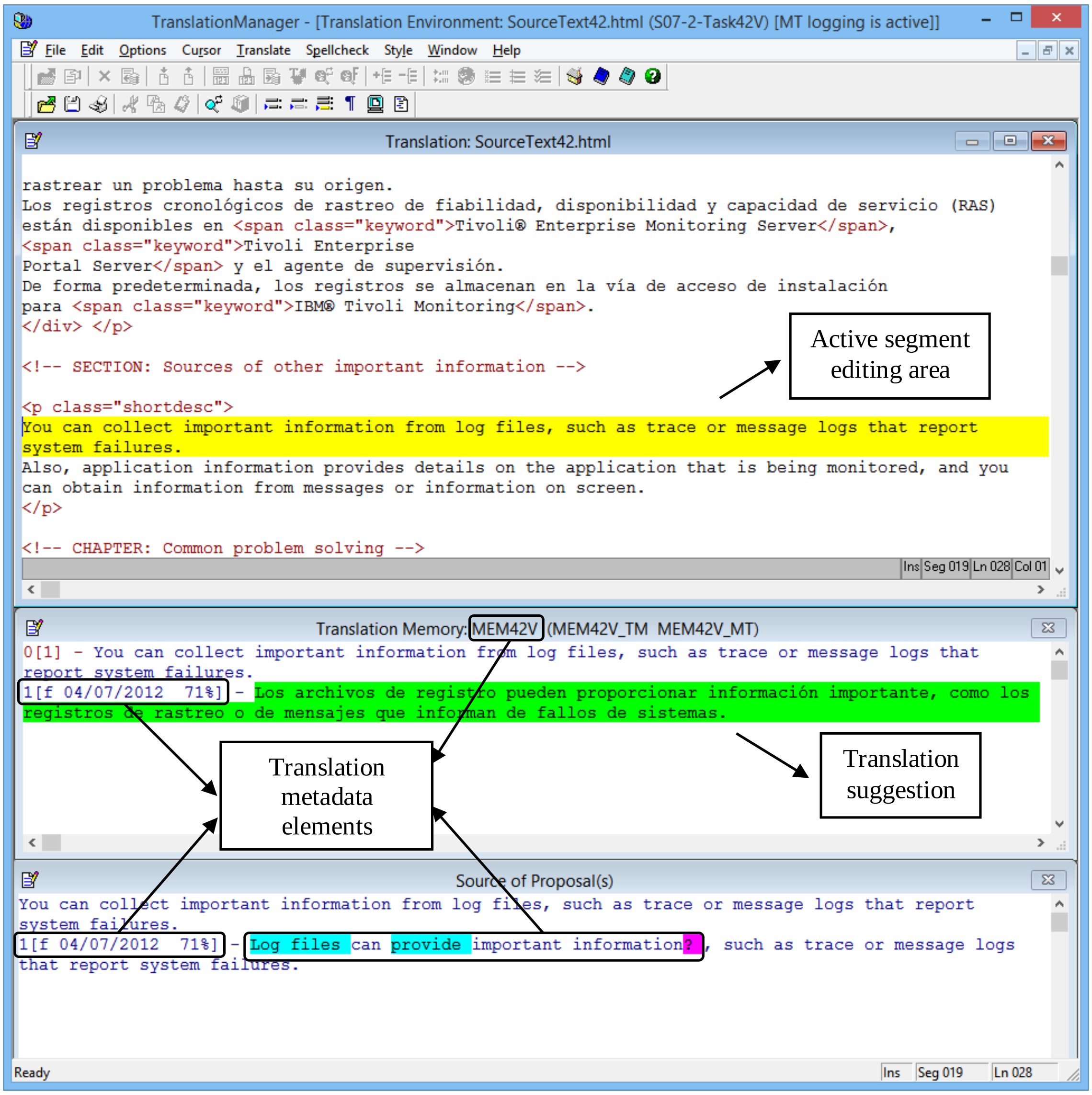
Translating from Scratch was always the first translation task, while the Visual task and the Blind task were performed in different orders depending on the participants. The two different source texts were used for the Visual and the Blind tasks and distributed evenly between the two tasks. Table 8 shows the distribution of task and text orders among the participants.

Table 8. Distribution of task and text orders among the participants

	1st Task		2nd Task		3rd Task	
Participant	Configuration	Text	Configuration	Text	Configuration	Text
P01	Scratch	10	Blind	31	Visual	42
P02	Scratch	10	Blind	42	Visual	31
P03	Scratch	10	Visual	42	Blind	31
P04	Scratch	10	Visual	42	Blind	31
P05	Scratch	10	Blind	42	Visual	31
P06	Scratch	10	Blind	42	Visual	31
P07	Scratch	10	Blind	31	Visual	42
P08	Scratch	10	Blind	31	Visual	42
P09	Scratch	10	Visual	31	Blind	42
P10	Scratch	10	Visual	31	Blind	42

In the Visual task, one translation suggestion was provided for each segment, and the translators had to actively insert it in the editing area and to edit it if they considered it to be a usable suggestion, or they could type their translation either from scratch or on top of the source text. The most common way for the translators to insert translation suggestions was by using a keyboard shortcut, although in some cases they preferred to copy and paste either the whole or parts of the suggestions from the Translation Memory pane into the active segment, combining the mouse and keyboard. In this task, translation suggestions were provided with metadata. The way IBM TranslationManager indicates this information is by placing a letter to the left of the suggestion: blank for exact matches, “f” for fuzzy matches and “m” for machine translation feeds. Additionally, in the case of fuzzy matches, the tool indicates the percentage of similarity and highlights the text portions that differ between the source text in the active segment and the source segment in the translation memory. Figure 1 illustrates how the Visual task was seen by translators in IBM TranslationManager.

Figure 1. The Visual task in IBM TranslationManager, with an indication of translation metadata.


TranslationManager - [Translation Environment: SourceText42.html (S07-2-Task42V) [MT logging is active]]
File
Edit
Options
Cursor
Translate
Spellcheck
Style
Window
Help
Translation: SourceText42.html
Active segment editing area
Ins
Seg 019
Ln 028
Col 01
Translation Memory: MEM42V (MEM42V_TM MEM42V_MT)
Translation metadata elements
Translation suggestion
Source of Proposal(s)
Ready
Ins
Seg 019
Ln 028

```
rastrear un problema hasta su origen.
Los registros cronológicos de rastreo de fiabilidad, disponibilidad y capacidad de servicio (RAS)
están disponibles en <span class="keyword">Tivoli® Enterprise Monitoring Server</span>,
<span class="keyword">Tivoli Enterprise
Portal Server</span> y el agente de supervisión.
De forma predeterminada, los registros se almacenan en la vía de acceso de instalación
para <span class="keyword">IBM® Tivoli Monitoring</span>.
</div> </p>

<!-- SECTION: Sources of other important information -->

<p class="shortdesc">
You can collect important information from log files, such as trace or message logs that report
system failures.
Also, application information provides details on the application that is being monitored, and you
can obtain information from messages or information on screen.
</p>

<!-- CHAPTER: Common problem solving -->
```

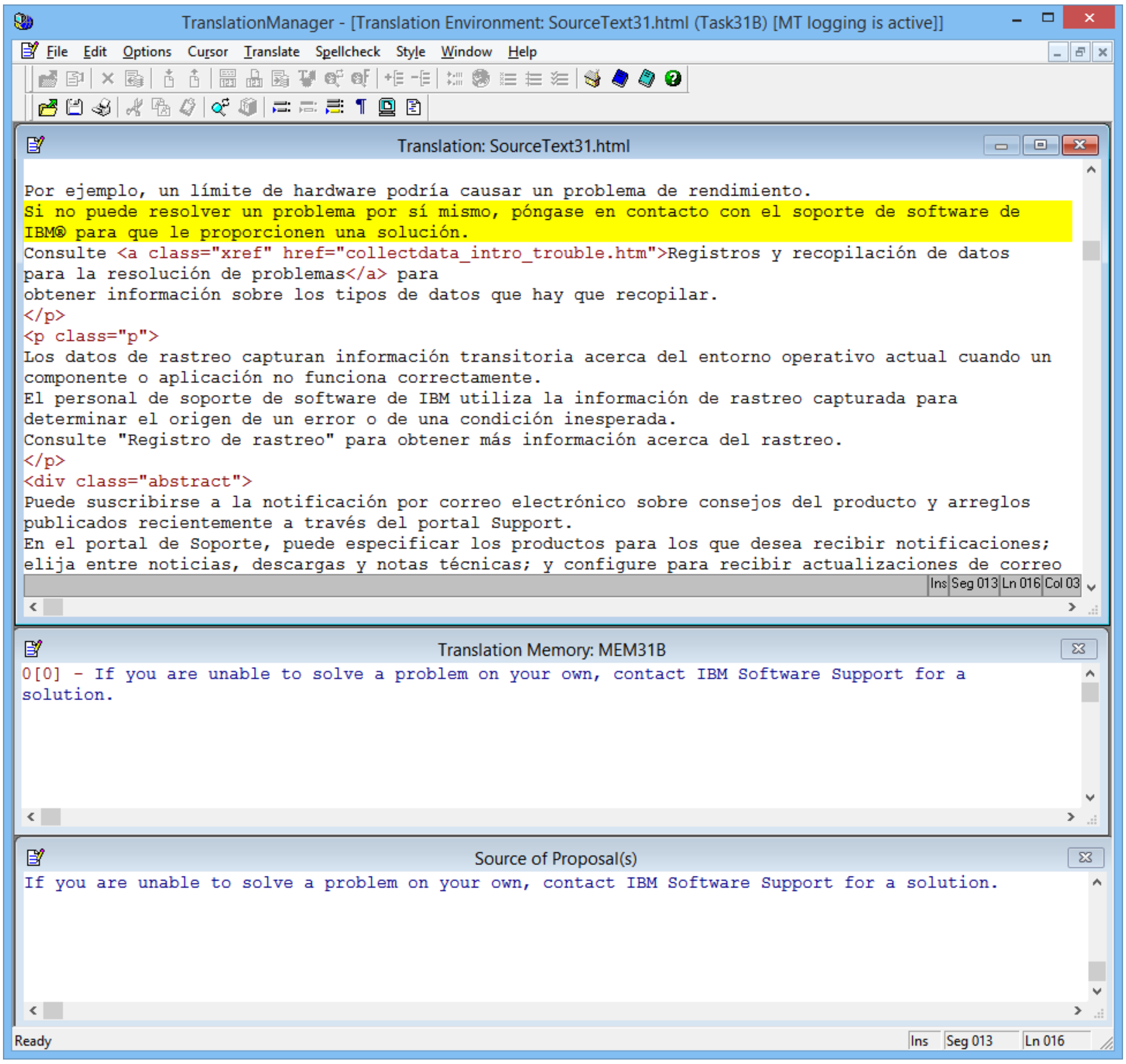
```
0[1] - You can collect important information from log files, such as trace or message logs that
report system failures.
1[f 04/07/2012  71%] - Los archivos de registro pueden proporcionar información importante, como los
registros de rastreo o de mensajes que informan de fallos de sistemas.
```

```
You can collect important information from log files, such as trace or message logs that report
system failures.
1[f 04/07/2012  71%] - Log files can provide important information?, such as trace or message logs
that report system failures.
```

Note: Metadata elements include: translation memory name; suggestion number (1); type of suggestion (“f” stands for “fuzzy”, in the example); date of last usage; match percentage; and differing text portions.

In the Blind task, there was also one translation suggestion per segment, but the suggestion had been previously inserted in the segment, so the file displayed as pre-translated text to be edited, instead of source text to be replaced with a translation suggestion. The application panes where the translation suggestions are usually displayed were empty, so no translation metadata were displayed. Figure 2 illustrates the Blind task in IBM TranslationManager.

Figure 2. The Blind task in IBM TranslationManager, with no translation metadata available


TranslationManager - [Translation Environment: SourceText31.html (Task31B) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText31.html
Ins Seg 013 Ln 016 Col 03
Translation Memory: MEM31B
Source of Proposal(s)
Ready
Ins Seg 013 Ln 016

```
Por ejemplo, un límite de hardware podría causar un problema de rendimiento.
Si no puede resolver un problema por sí mismo, póngase en contacto con el soporte de software de
IBM® para que le proporcionen una solución.
Consulte <a class="xref" href="collectdata_intro_trouble.htm">Registros y recopilación de datos
para la resolución de problemas</a> para
obtener información sobre los tipos de datos que hay que recopilar.
</p>
<p class="p">
Los datos de rastreo capturan información transitoria acerca del entorno operativo actual cuando un
componente o aplicación no funciona correctamente.
El personal de soporte de software de IBM utiliza la información de rastreo capturada para
determinar el origen de un error o de una condición inesperada.
Consulte "Registro de rastreo" para obtener más información acerca del rastreo.
</p>
<div class="abstract">
Puede suscribirse a la notificación por correo electrónico sobre consejos del producto y arreglos
publicados recientemente a través del portal Support.
En el portal de Soporte, puede especificar los productos para los que desea recibir notificaciones;
elija entre noticias, descargas y notas técnicas; y configure para recibir actualizaciones de correo
```

```
0[0] - If you are unable to solve a problem on your own, contact IBM Software Support for a
solution.
```

```
If you are unable to solve a problem on your own, contact IBM Software Support for a solution.
```

Participants were allowed to add any glossaries and to consult any references in all of the tasks, although some of them assumed that they were not allowed to. In any case, those that used glossaries in one of the tasks also used them in the other tasks.

4.5.7.2. Text presentation

As mentioned in previous sections, all translators started the experiment by copying SourceText_Copy into a text editor. Then they translated SourceText10 from scratch in IBM TM/2. The two main translation tasks came immediately after this, when the translators were asked to translate either SourceText31 or SourceText42 in either the

Visual or the Blind task, according to Table 8 (page 56). All the start texts were presented in printed form and formatted like the original manual, as can be seen in Appendices 1 to 4. The titles were highlighted with a text marker as they were not to be translated. For the translation tasks (Scratch, Visual and Blind) the translatable source text also displayed within the translation tool.

4.5.7.3. Instructions

All the communication between the researcher and the participants was done in Spanish, except for one participant, who occasionally preferred to communicate in Catalan. The instructions for the Copy and the Scratch tasks were given orally. The instructions for the Visual and the Blind tasks were given in printed form (see Appendix 7 and Appendix 8) first and then also explained orally, based on any questions asked by the translators. The instructions for the Visual task mentioned that the TM/2 folder (project) to be translated contained two translation memories: one that was taken from a previous release of the same product and another containing machine-translated segments from the same MT engine they had been using for IBM projects in the company, tuned for the specific product family. The instructions for the Blind task were mostly identical, the only difference being that it stated the TM/2 folder contained no translation memories and the segments had been pre-translated with translation suggestions from the same types of sources as in the Visual task (an IBM translation memory and the customised MT engine). Due to a slip of my own, there was also an unintended difference in the last paragraph of the instructions: the instructions for the Visual task stated that the participants were supposed to "traducir" (translate) the text they had been sent, whereas the instructions for the Blind task stated that they were supposed to "revisar" (revise or proofread) the text.[9]

The instructions also mentioned that the participants' final translations were going to be assessed by the company's reviewers afterwards, as in a normal IBM project. No specific instructions were given on the quality expectations; participants were told that they should produce their translations with the same quality level of their regular IBM assignments. Since those assignments (and the tasks in the experiment) presented MT suggestions in a traditional TM workflow, no specific post-editing guidelines were given. In practice, this corresponds to what is known as "light revision" and "light post-editing".

[9] My sincere thanks to Arnt Lykke Jakobsen for bringing up this issue during the thesis defence.

Before they started, I also explained in general terms the main purposes of the experiment. No time limit was set for any of the tasks. The interval between the tasks was only the time necessary to present the instructions for the following task, locate the folder to be opened by the translators and prepare the research equipment to record the following task. This interval was not timed, but ranged roughly from two to five minutes.

4.6. Data collection methods

The current study focus on what happens *while* translation takes place, i.e. *while* translators translate, and as such lies within what is known as translation process research. This is also a workplace study, as it investigates how translators work in their normal work environment, rather than having translators come to a lab prepared for an experiment. Workplace studies have the advantage of investigating the translation activity in a more authentic setting, but they also cause increased difficulty to the researcher, as it is harder to control the experiment conditions.

In what follows I will detail each of the methods used in the current study for collecting the data. Section 4.8 will explain how these data are processed and analysed.

4.6.1. Keystroke logging

Keystroke logging consists in tracking all the keyboard actions performed by the translators while they work on a given translation task. This method allows the researcher to perform several kinds of analysis, the most prominent of them being pause analysis (Jakobsen 2003; Dragsted 2004; O’Brien 2006b). In the current study, keystroke-logging tools are used to measure the amount of time and the number of keystrokes used by the translators to produce a given translated segment.

IBM TranslationManager, the translation memory tool used in the experiment, comes with an integrated command-line tool called MTeval, which can perform post-task analyses based on the exported folders (projects). Because it works in conjunction with IBM TranslationManager, MTeval provides the logging data divided by text segments for easier analysis. It provides information on the time taken to process each segment, the number of characters typed, the kind of translation suggestion offered by the translation tool and the kind of suggestion actually used by the translator.

MTeval is used by IBM and its translation vendors to assess post-editing effort and to evaluate the quality of the MT engines being used. For the purposes of the experiment, it was used to collect information on time and typing effort, as a complement to other more complex methods of data collection. One limitation MTeval has for my intended purposes is that it resets the time counter whenever a translation suggestion is inserted in the segment. For example, when a segment is activated, the counter starts and, if the translator edits on top of the existing text, MTeval will count the time correctly until the translator closes the segment and moves on to the next segment. This poses no problem when the existing text in the active segment is a pre-translation and no further suggested translations are available, as is the case in the Blind task. However, in the Visual task, the active segment will initially contain a copy of the source text and the translator will eventually replace it with the translation suggestion. As soon as this happens, the time counter is reset and we lose track of the time spent between the moment when the segment was activated (opened) and the moment when the translation suggested was inserted. The time that is discarded is relevant for my research interests, as it represents the time it takes translators to make their decisions about the translation suggestions, and precisely the period during which they might look at the translation metadata.

For this reason, I decided to look for an additional keystroke logging tool. The first such tool I considered was Translog.[10] It is well documented and has been mentioned in several academic papers in translation process research (Rydning 2002; Jakobsen 2006; Göpferich and Alves et al. 2009), but it was designed to track only those keyboard and mouse activities that take place within Translog itself. Since my experiment required translators to work with a translation memory tool, Translog was not a viable option.

After considering other alternatives[11], I opted for Inputlog[12] (Leijten and van Waes 2013). It is a lightweight and stable application that offers system-wide logging capabilities. This, in practice, allows one to track any application running on the operating system, including the translation tool used in my experiment. In addition to providing detailed information about keyboard and mouse activity, it logs system events, can integrate eye-tracking data and includes several types of analyses (reports).[13] It provides XML files as the output for the different types of analyses, which can then be processed

[10] http://www.translog.dk

[11] A comparison of several logging programmes can be found at http://www.writingpro.eu/logging_programs.php

[12] http://www.inputlog.net

[13] The program version I used (5.0.1.26) offered the following types of analyses: general, summary, pause, linear, focus, S-notation and W-notation. Inputlog version 6.0, the latest release by the time of writing, has added several new types of analysis, including the analysis of Translog files.

in external tools, such as Excel. Inputlog also includes a special replay function module, but this is only available when translation tasks are performed in Microsoft Word, so I was not able to use this feature. As an added benefit, Inputlog is free of charge for research purposes.

4.6.2. Screen recording

In order to visualise what the translators were doing at any given moment, I used a screen-recording tool. By recording and then watching the activity of each participant translator, I was able to view how they dealt with each particular segment.

Several studies have reported on the use of Camtasia Studio[14] in translation process research (O'Brien 2006a; Buchweitz and Alves 2006; Göpferich and Alves et al. 2009; Angelone 2010). I preferred to use BB FlashBack (Enríquez Raído 2011; Morado Vázquez 2012) instead, as it seemed more intuitive and offered a free version (BB FlashBack Express) that included all the features I needed for analysis. It allows recording and viewing of all user activities, keystrokes, mouse clicks, ambient sounds and even facial expressions.[15] It is a very lightweight application in recording mode, and I was able to set it up easily and quickly on the participants' computers, even with different system specifications.

In the pilot experiment, BB FlashBack was used to manually calculate the time and the number of keystrokes for individual segments (see section 4.4). In the main experiment, it was used mainly to help understand in a visual way what the translators had done and to identify extraneous reasons for pauses. Its face-recording feature can help determine whether a pause was caused by extra attention to the text or happened simply because the participant was distracted, took a drink of water, spoke with someone, etc. BB FlashBack was also used as a back-up method – in case something went wrong with Inputlog and/or MTeval – and for checking specific unclear situations, such as when the transition between segments could not be properly identified from the keystroke logs.

[14] http://www.techsmith.com/camtasia.html (A free trial version is available for 30 days.)

[15] BB FlashBack is also available in Standard and Pro versions. These pay versions have extra features such as adding text, sound and images to a movie; editing the movie and adding effects; exporting the movie to different video formats; online sharing to custom servers. More information on each version and a comparison chart is available at http://www.bbsoftware.co.uk/BBFlashBack/CompareEditions.aspx

4.6.3. Eye tracking

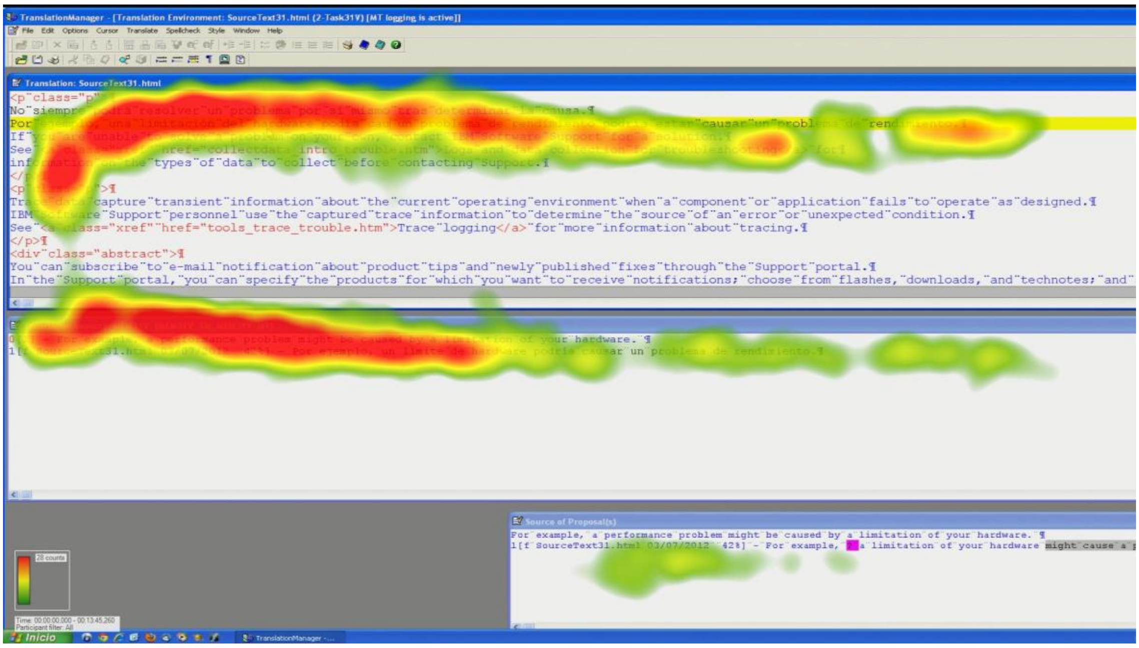
Eye tracking makes it possible to identify where on the screen a translator is looking at during a translation task. The decision to integrate this technique into my study came from the difficulty I found in the pilot experiment of identifying the segment translators were processing at a given moment, especially in the self-revising phase. It was also expected that eye tracking would help corroborate my assumption that translation metadata are relevant for translators.

The usefulness of eye tracking for translation research has been confirmed by studies such as O'Brien (2006a), and is based on Just and Carpenter's (1980) eye-mind assumption that "there is no appreciable lag between what is being fixated and what is being processed" (1980: 331).

An eye tracker detects eye movements and maps them onto what is displayed on a screen or virtually any other surface. It can also collect pupillometric data, i.e. information on the size of the eye's pupil at a given moment, as an indicator of stress. The data collected by the eye tracker contains, for a given point in time, the (x, y) coordinates of the surface that each eye was looking at and the pupil diameter of each eye. The frequency setting on the eye tracker determines how many points in time are captured. For example, an eye tracker working at 120 Hz produces 120 data samples in one second, or one sample every 8.3 milliseconds. This basic data can be complemented with the z position of the eyes (gaze depth) and indicators of confidence for the recorded values. It can also include non-eye related data such as keyboard and mouse events. Depending on the eye-tracking software used, it is then possible to run several kinds of analyses with the recorded data in combination with what was being displayed on the screen.

The two most common types of visualisations for eye-tracking data are heat maps and gaze plots. Heat maps display the gaze data as a graphical representation where all eye fixations over the length of a recording – or part of it – are combined and represented as coloured areas (Figure 3). The researcher can define either the number or the duration of fixations used for drawing the heat maps. The result can be visualised either dynamically, by overlapping the cumulative fixations to the recorded video image, or statically, by overlapping all fixations for a given interval over a fixed image that represents what was seen on the screen during the interval. Either way, by analysing heat maps the researcher can tell qualitatively which areas of the screen where looked at most frequently (fixation count) or for longer periods (fixation duration).

Figure 3. A heat map (example from the Visual task in my experiment)

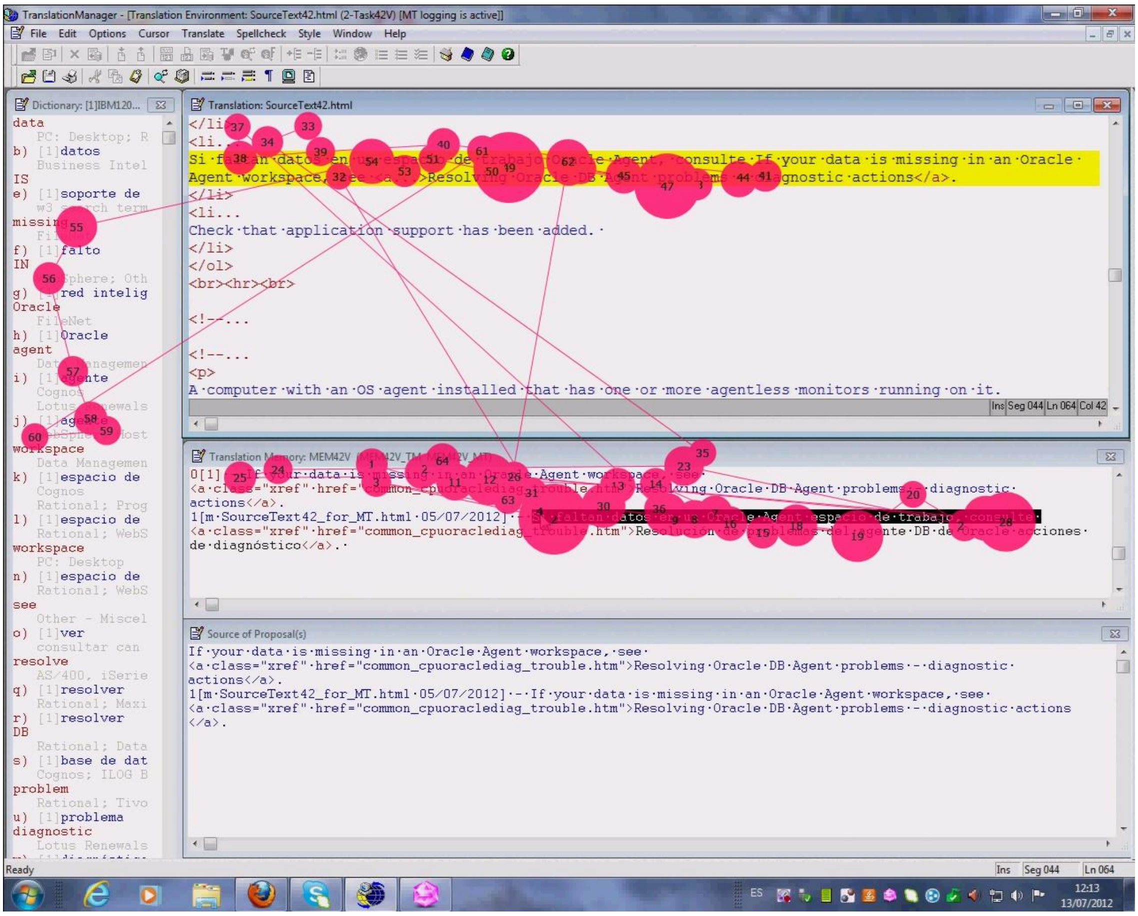

TranslationManager - [Translation Environment: SourceText31.html (2-Task31V) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText31.html
Source of Proposal(s)
28 counts
Time: 00:00:00.000 - 00:13:45.260
Participant filter: All
Inicio
TranslationManager - ...

```
<p"class="p"[illegible]
No"siempre [illegible] resolver"un"problema"por"sí"mismo [illegible] la"causa.¶
Por [illegible] estar"causar"un"problema"de"rend[illegible]
If [illegible] unable [illegible] IBM"Software"Support [illegible]
See [illegible] href="collectdata_intro_trouble.htm"> [illegible]
information [illegible] the"types"of"data"to"collect"before"contacting"Support.¶
</p[illegible]
<p[illegible]>¶
Tra[illegible]capture"transient"information"about"the"current"operating"environment"when"a"component"or"application"fails"to"operate"as"designed.¶
IBM[illegible]are"Support"personnel"use"the"captured"trace"information"to"determine"the"source"of"an"error"or"unexpected"condition.¶
See"<a class="xref""href="tools_trace_trouble.htm">Trace"logging</a>"for"more"information"about"tracing.¶
</p>¶
<div"class="abstract">¶
You"can"subscribe"to"e-mail"notification"about"product"tips"and"newly"published"fixes"through"the"Support"portal.¶
In"the"Support"portal,"you"can"specify"the"products"for"which"you"want"to"receive"notifications;"choose"from"flashes,"downloads,"and"technotes;"and"
```

```
0[[illegible] performance problem might be caused by a limitation of your"hardware."¶
1[[illegible]31.html [illegible] Por ejemplo, un límite de hardware podría causar un problema de rendimiento.¶
```

```
For"example,"a"performance"problem"might"be"caused"by"a"limitation"of"your"hardware."¶
1[f"SourceText31.html"03/07/2012""42%]"-"For"example,"[illegible]"a"limitation"of"your"hardware"might"cause"a"p
```

Figure 4. A gaze plot (example from the Visual task in my experiment)


TranslationManager - [Translation Environment: SourceText42.html (2-Task42V) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Dictionary: [1]IBM120...
Translation: SourceText42.html
Translation Memory: MEM42V (MEM42V_TM MEM42V_MT)
Source of Proposal(s)
Ins Seg 044 Ln 064 Col 42
Ready
Ins Seg 044 Ln 064
ES
12:13
13/07/2012
1
2
3
4
7
8
9
11
12
13
14
15
16
18
19
20
23
24
25
26
28
30
31
32
33
34
35
36
37
38
39
40
41
44
45
47
49
50
51
53
54
55
56
57
58
59
60
61
62
63
64

```
data
    PC: Desktop; R
b) [1]datos
    Business Intel
IS
e) [1]soporte de
    w3 search term
missing
    Fi[illegible]
f) [1]falto
IN
    [illegible]Sphere; Oth
g) [1]red intelig
Oracle
    FileNet
h) [1]Oracle
agent
    Data Managemen
i) [1]agente
    Cognos
    Lotus [illegible]ewals
j) [1]ag[illegible]
    [illegible]bSphe[illegible]ost
workspace
    Data Managemen
k) [1]espacio de
    Cognos
    Rational; Prog
l) [1]espacio de
    Rational; WebS
workspace
    PC: Desktop
n) [1]espacio de
    Rational; WebS
see
    Other - Miscel
o) [1]ver
    consultar can
resolve
    AS/400, iSerie
q) [1]resolver
    Rational; Maxi
r) [1]resolver
DB
    Rational; Data
s) [1]base de dat
    Cognos; ILOG B
problem
    Rational; Tivo
u) [1]problema
diagnostic
    Lotus Renewals
```

```
</li>
<li...
Si·f[illegible]an·datos·en·un·espacio·de·trabajo·Oracle·Agent,·consulte·If·your·data·is·missing·in·an·Oracle·
Agent·workspace,·see·<a...>Resolving·Oracle·DB·Agent·problems·-·diagnostic·actions</a>.
</li>
<li...
Check·that·application·support·has·been·added.·
</li>
</ol>
<br><hr><br>

<!--...

<!--...
<p>
A·computer·with·an·OS·agent·installed·that·has·one·or·more·agentless·monitors·running·on·it.
```

```
0[1]·-·If·your·data·is·missing·in·an·Oracle·Agent·workspace,·see·
<a·class="xref"·href="common_cpuoraclediag_trouble.htm">Resolving·Oracle·DB·Agent·problems·-·diagnostic·
actions</a>.
1[m·SourceText42_for_MT.html·05/07/2012]·-·Si·faltan·datos·en·un·Oracle·Agent·espacio·de·trabajo,·consulte·
<a·class="xref"·href="common_cpuoraclediag_trouble.htm">Resolución·de·problemas·del·agente·DB·de·Oracle·acciones·
de·diagnóstico</a>.·
```

```
If·your·data·is·missing·in·an·Oracle·Agent·workspace,·see·
<a·class="xref"·href="common_cpuoraclediag_trouble.htm">Resolving·Oracle·DB·Agent·problems·-·diagnostic·
actions</a>.
1[m·SourceText42_for_MT.html·05/07/2012]·-·If·your·data·is·missing·in·an·Oracle·Agent·workspace,·see·
<a·class="xref"·href="common_cpuoraclediag_trouble.htm">Resolving·Oracle·DB·Agent·problems·-·diagnostic·actions
</a>.
```

Gaze plots are a different way of displaying data on gaze behaviour (Figure 4). In this case, fixations are represented by numbered circles, where the radius of each circle is a function of the fixation duration, and the numbers represent the order in which the fixations occurred. Straight lines connecting the circles in a gaze plot represent the path followed by the eyes (saccades) between contiguous fixations. As is the case with heat maps, gaze plots can also be visualised as static or dynamic images.

There are several types of eye trackers available. In translation process research, the most common types use an internal video camera to capture infrared light reflected on the cornea to record eye movements. These camera-based eye trackers can be head-mounted (usually in the form of helmets or glasses) or desktop-mounted. Head-mounted eye trackers tend to be more accurate and have the advantage of being able to follow the eye-movements even with large head movements. Their disadvantage is their intrusiveness, which has a strong negative impact on the ecological validity of any experiment.

Desktop-mounted eye trackers can rely on some kind of head support (chin rest, forehead rest or bite bar) or they can allow for free head movement. Eye trackers that rely on head support are mostly unsuitable for any research where participants have to use a keyboard. For translation process experiments, we are thus left with desktop-mounted eye trackers with no head support. Even so, there is a choice between eye trackers that are built into a screen monitor and those that come as stand-alone units. The types of eye tracker most often used in translation process research and mentioned in the articles published in the field are desktop-mounted and built into a screen monitor.

O'Brien (2009a) discusses several aspects that need to be taken into account when designing an experiment that includes eye tracking. In particular, she formulates the Translation Studies version of "the observer's paradox" (Labov 1972):

> [W]e wish to observe what professional translators "normally" do, but we remove them from their "normal" work environments in order to do so. The fact that the eye-tracking monitor is most likely different in shape and size from their usual monitor, or that the operating system, software, version numbers, language packs, screen layout or even keyboard type differ from their usual work environment may have an impact on their performance. The research community should not abandon research because of these challenges, but, where a "normal work environment" is an important factor, the eye-tracking environment ought to be set up in such a way that the participant is familiar and comfortable with it. (O'Brien 2009a: 253–254)

Since I wanted to have participants work in a scenario as close as possible to their regular work environment, I thought the best option would be to use a stand-alone eye tracker. Based on my recommendation, my research group at the URV chose to purchase a Tobii X120 among all the options from competing brands. This model can be adapted to virtually any screen size and was suited for the large, panoramic screen monitors used by my participants, which ranged from 19” to 23” with resolutions from 1280 × 1024 to 1900 × 1080 pixels. The eye-tracking software provided with Tobii eye trackers is called Tobii Studio.

Because I wanted to investigate the behaviour of translators in their workplace, I adapted my expectations concerning the quality of the eye-tracking data I would be able to obtain. For example, I discarded analysing pupil dilation, as this would have required controlling my experiment environment for technical factors such as lighting, sound and vibration (O’Brien 2009a; Hvelplund 2011: 103) as well as for “human” factors such as eye colour, eye make-up or whether participants were allowed to have coffee before or during the experiment. As it happened, the environment where my translations took place was a very noisy office, with variable light conditions, and translators drank coffee or tea several times a day. Yet it was the environment where the participants were used to working every day. Asking them to work in a quiet room might have affected the validity of the experiment more than the distracting factors might have affected the eye-tracking data. If a researcher were interested in pupil-dilation data, it would be necessary to find a different balance between validity and accuracy.

Even disregarding the difficulties related to pupil dilation, eye tracking is perhaps the most challenging of the methods I used for data collection, both when running the experiment and during data processing. I was faced with other problems such as the installation and calibration of equipment. The kind of eye tracker I used, and the way I designed my experiment, required that I measure several sizes, distances and angles between the eye tracker, the computer screen and the table, for each translator. These measurements were made manually with the rudimentary tools provided by the eye-tracker manufacturer, then introduced into the eye-tracker configuration software. Perhaps because of the insufficient accuracy of the measurement methods or other factors, sometimes it was not possible to obtain an optimal calibration even after several rounds. This might have been improved with repeated attempts at calibration, but with a concomitant loss of ecological validity, as my translators stood around waiting to perform and wondering what was wrong. In this, as in much else, there was a trade-off between accuracy and ecology.

The second major complication was the need to use a video-capture device to get the image from the translator's computer screen and map it with the eye-tracking data captured by Tobii Studio on the researcher's laptop. The first difficulty arose when selecting a video codec to use. Only the *Microsoft Video 1* codec provided satisfactory results – all my attempts with other codecs recommended by Tobii, including TechSmith, Morgan MJPEG and Xvid, resulted in either a blank screen image being captured by Tobii Studio or in extremely large files being created on the hard disk.

Even with the functional codec, a major challenge was some random behaviour with the image coming from the translator's computer to the researcher's computer, where Tobii Studio was running. In some cases, after a period of correct capturing, the image became whitish and remained so until the end of the recording. This happened both in Tobii Studio and in Epiphan Capture Tool (the software tool that came with the video capture device), which suggests that the problem was not related to the eye-tracking software – it may have been due to the video capture device, the laptop I was using or the connecting cables and adapters. Although the resulting images were not ideal, in most cases they still allowed me to identify the relevant areas of the screen where translators were working.

The most problematic issue was that part of the screen image coming from the translator's computer was not captured, for some of the participants. Comparing the image of the actual computer screen as captured by BB FlashBack with the corresponding image as captured by Tobii Studio through the video capture device, it is noticeable that a wide vertical strip of image was lost on the right or left side of the screen. An example of such a case is presented in Figure 38 in section 5.5.2. Again, this behaviour was also captured by Epiphan Capture Tool. Even after trying all possible settings in the video capture tool, installing and reinstalling hardware and software and checking with Alt64/Tobii support, I was not able to solve the problem and had to live with it. As a result, in some cases the eye-tracking data does not map correctly to the screen image. As it happened, this can often be compensated for by the fact that my areas of interest in this particular study are "horizontal" and somewhat spread apart, while the part that was cut from the translator's screen was systematically vertical. Since I was mainly interested in detecting which screen element the translators were focusing at a particular moment, it still seemed possible to proceed. This would not be the case in many other experiments.

In the end, however, due to gaze data loss, unsatisfactory calibration and incomplete screen image coming through the video capture device, the use of eye-tracking data as I had planned initially – a quantitative analysis across all subjects by analysing the gaze duration according to three main areas of interest – was not possible. Nevertheless, for certain subjects and during certain parts of the recordings, I have been able to make a qualitative analysis of the translators' behaviours. For example, I have been able to observe how often they shifted their attention between the target text, the source text and the translation suggestions or what editing strategies they used according to the different suggestion types. Examples of such analysis are presented in section 5.5.

The more complex the technology, the higher the probability that such problems will arise, and the methods of data analysis may have to be adapted accordingly.

4.6.4. Interviews and retrospection

In order to try to understand what goes on in the translators' minds while they translate, several methods have been used in our field. One example is think-aloud protocols (TAPs), which have the translators "think out loud" while they are translating, i.e. verbalising their problems and problem-solving processes. This is presumed to indicate the normal thought processes of translators, in line with the assumptions made by Ericsson and Simon (1984) in psychology research. TAPs have been used in process research for a couple of decades, with interesting results (Kussmaul and Tirkkonen-Condit 1995; Bernardini 1999; Rydning 2002; Jääskeläinen 2002; Hansen 2005). There remain doubts, however, as to their ecological validity. After all, translators do not normally talk out loud while they are working. Further studies indicate that simultaneous TAPs (speaking while translating) actually slow down the translation process (Krings 2001; Jakobsen 2003), which makes them unsuitable in cases where one wants to measure translation speed, as was the case in my experiment.

Some studies suggest that it is better to have a screen recording of a normal (i.e. with no TAP) translation performance, and then have the translator comment on the recording as it is played back afterwards. Such retrospective methods would be more ecologically valid, although they also give the subject translators ample scope for self-justification, constructing performance narratives after the event. If that caveat is accepted, several variations are possible.

In my experiment, I used post-performance interviews (dialogues), combined with retrospection with replay (Hansen 2006; 2008). This combination of methods does not interfere with the translation process and should provide the “retrieval cues” that are necessary to stimulate translators to comment on the task just performed (Hansen 2008: 4). This can provide information not only on translator’s feelings and task satisfaction, but also on the pertinent translation norms, since subjects tend to express what they thought they *should* be doing. It was also hoped that both kinds of interviews would provide information on the translators’ perception regarding the role of metadata in the different tasks.

The interviews were first carried out as semi-structured dialogues, right after all the translation tasks had been completed. Each translator was interviewed in Spanish[16] and answered the following main questions:

1) Do you think you translated *faster* in any of the environments? If so, in which one?
2) Do you think the quality of your final translation was *better* in any of them? If so, in which one?
3) In which environment did you feel more *comfortable* working?

The retrospection with replay took place immediately after the dialogues: the translators were invited to watch selected passages of their performance recordings together with the researcher and to answer specific questions about certain aspects of the translation tasks. For this phase, I used the RTA (Retrospective Think Aloud) feature in Tobii Studio, which allows the image and voice of the translators to be recorded while they watch the recording of their translation process combined with visual eye-tracking information.

4.6.5. Translation reviews

Quality is one of the most complex aspects to assess in translation-process research. Different methods have been used in the literature, most of them based on human revision, such as applying the LISA QA grid or other types of purpose-built evaluation grids, as well as revision time. The more recent TAUS Dynamic Quality Framework (DQF) could also provide a framework for similar kinds of human assessment. For my study I went

[16] Except for one participant, who preferred to talk in Catalan.

along the same lines and chose to use a method of human revision, preferably as close as possible to the “real world” of my translators.

The preliminary interviews I had with different people in MSS indicated that for the kind of material used in the experiment, the translated files did not always undergo systematic revision by a second professional, but they were always spot-checked by the corresponding project manager following certain guidelines. Some problems in the translation (tags, inconsistency, spelling) were considered more serious than other problems (style), even though the latter could be deemed serious in other contexts.

Based on these findings, two of the company reviewers were asked to assess the translations in a way as close as possible to what they would do in a normal project. The only difference was that they had to review the translations in Word and each segment at once across all ten translators, rather than reviewing the whole files for each of the translators in IBM TranslationManager. To make sure reviewers were not trapped by recurrent mistakes made by the translators, I added instructions or comments in specific segments. Additionally, I asked the reviewers to go back participant by participant and look for inconsistencies in specific terms (those we knew were recurrent in the source texts).

One of the reviewers was a 38 year-old man with 12 years of experience revising IBM material, who worked in-house and was also a project manager and technical support. The other was a 46 year-old woman with 18 years of experience revising IBM material who worked from home as a freelancer. As a research method, the use of professional reviewers is ecologically sound but possibly expensive. Fortunately, my sponsoring project provided the funding necessary to pay these professionals.

4.7. Equipment and software

In view of the data collection needs described above, my experiment design required me to install and operate tools for keyboard and mouse logging, screen recording, face and voice recording, and eye tracking.

At first I considered installing all the data collection tools on each translator’s computer prior to the experiment. However, there were risks associated with running the eye-tracking software with a new configuration each time and the need to make sure all the computers offered the minimum system requirements for Tobii Studio. For these reasons, I opted to run Tobii Studio always from my own computer and install only the other tools on the translator's computers.

Figure 5. Photo illustrating the experimental set-up

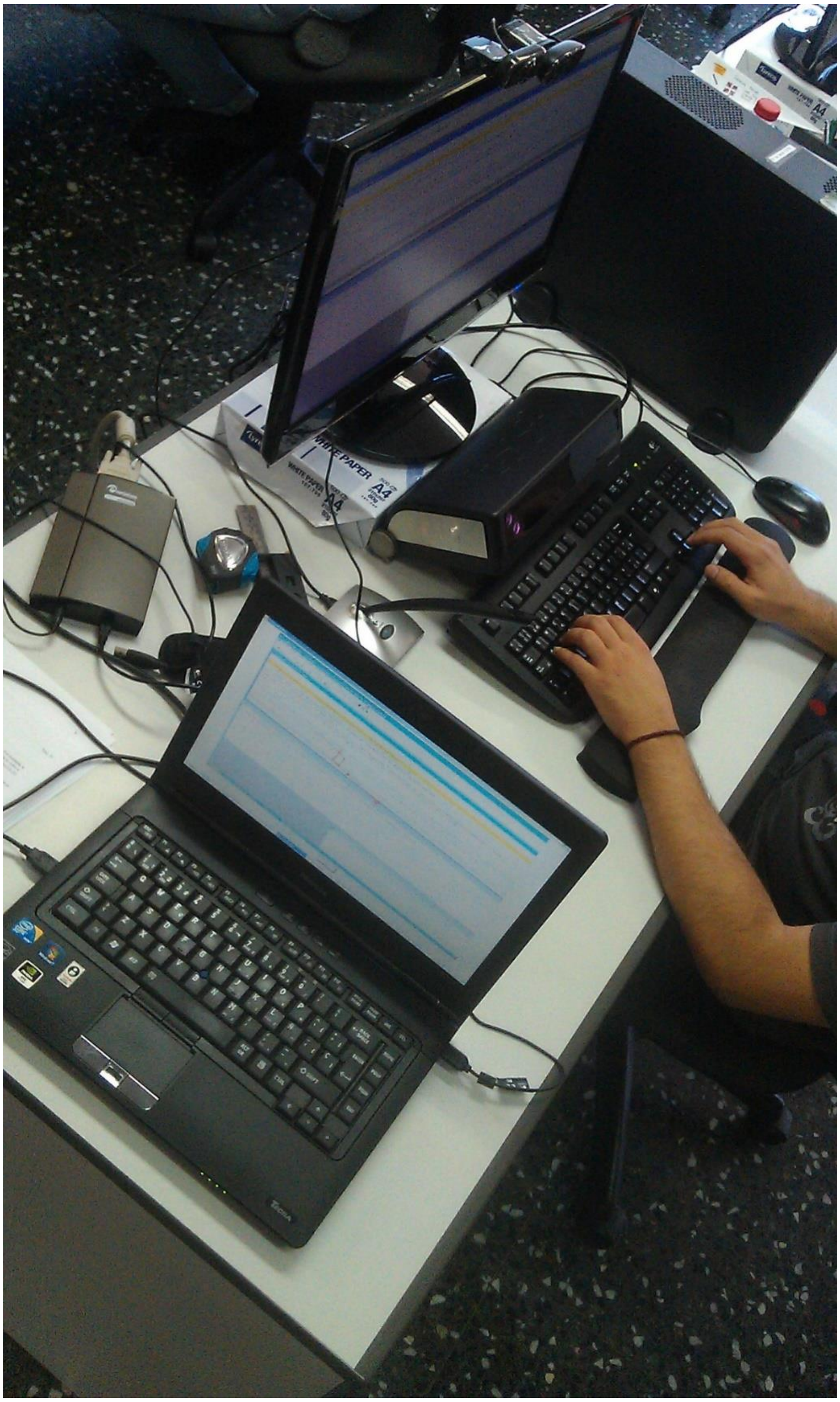
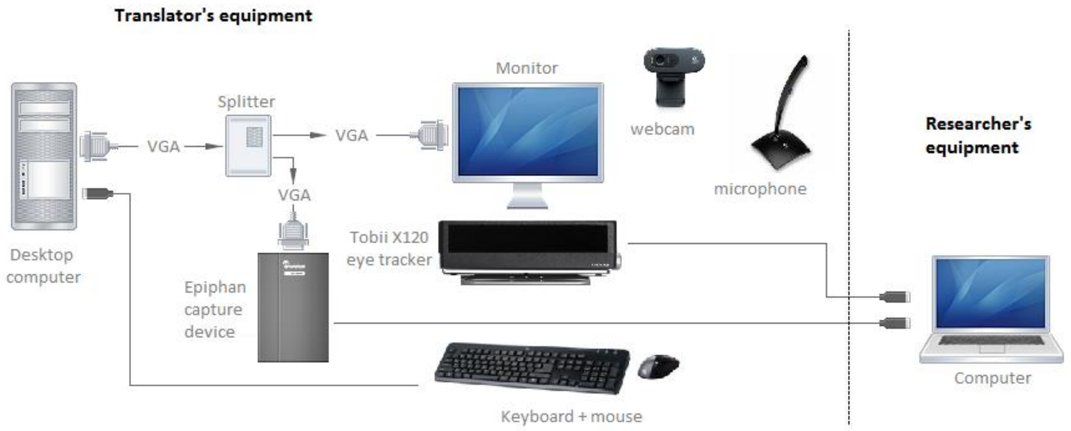


Figure 5 illustrates the set-up of the experiment. It shows a translator working on his computer and my computer at the side. The eye tracker was positioned between the screen and the keyboard. A webcam placed on the translator's monitor captured the image of the translator's face. A stand microphone placed on the table in front of the translator captured any conversations between the translator and the researcher, as well as any ambient sounds. BB FlashBack Pro 3 Recorder version 3.3.5.2273 was installed on the translator's computer and recorded the data captured by both the webcam and the microphone, as well as the full screen image, and the keyboard and mouse activity. Additionally, Inputlog version 5.0.1.26 was installed on the translator's computer and also recorded the keyboard and mouse activity.

At the same time, a video capture device (Epiphan DVI2USB) provided by the eye-tracker distributor Alt64/Tobii was connected through an image splitter to the video output of the translator's computer to send its full screen image to the researcher's computer. Finally, the eye tracker was also connected through a LAN cable to the researcher's computer. There, Tobii Studio 3.1.3 combined the eye-tracking data with the data from the video capture device using Tobii Studio's External Video option. It is worth noting that Tobii Studio could not track the translator's keyboard and mouse activity, as Tobii Studio was installed on the researcher's computer.

A simplified diagram of the experimental set-up is shown in Figure 6.

Figure 6. Simplified representation of the experimental set-up



4.8. Data analysis

4.8.1. Time and typing

For each segment, the time spent and the number of keystrokes were measured with Inputlog and MTeval. When noticeable differences were found between the values calculated with those tools, a second check was performed in order to understand the differences. An expected difference occurred in the Visual task because of the way MTeval calculates the active time in a segment, as explained in section 4.6.1 above. BBF provided visual guidance and helped confirm any dubious cases. For one participant (P02) there was no sound or image available from BBF due to technical problems, and the recordings in Tobii Studio were used instead.

Times and keystrokes were discounted when translators were checking external references, configuring the tools or talking. Times were not discounted when translators were checking terminology within the translation tool. Control keys such as Ctrl, Shift, Enter, Caps Lock and navigation arrows were not taken into account in the calculation of keystrokes. Other keys that have a direct impact on the number of characters produced were taken into account, such as Backspace, Delete and Space.

It is true that my method is a count of actions and does not take into account the number of linguistic transformations that take place or the mouse usage patterns. However, it takes into account every relevant action performed on the keyboard and seems to be a good compromise solution, if the analysis of linguistic transformations is to be avoided for the sake of simplifying the data analysis.

4.8.2. Quality

All translations were assessed for quality by two reviewers, who had been revising this type of material for many years (see 4.6.5). The reviewers revised the translations as Word documents, highlighting their corrections with the Track Changes feature. They were instructed to correct the errors that were deemed severe for this type of translation project and to ignore other types of errors. The severity of errors had been previously identified through a series of interviews with project managers in the company. Errors related to misinterpretation of the original, missing or added information, tag corruption and misspelt brand names scored two points. Errors such as inconsistencies, misspellings, wrong grammar and punctuation scored one point. Since translators were not instructed to follow any particular glossary, term consistency was only considered within and between the translations of each particular translator, not between their term choices and the terms in any IBM glossary. Other text issues such as those related to style and fluency were not taken into account. The researcher acted as a third reviewer, making decisions when the two reviewers had very different opinions and marking any obvious errors that had not been detected by the reviewers. These included:

- the inconsistent use of the initial article in the first sentences of the glossary entries (in the English source, a definite or indefinite article is always present, while in Spanish it is customary to omit the article);
- the inconsistent use of the registered trade mark symbol ® (the source was inconsistent, but translators were expected to be consistent in their decisions of either using the symbol, not using the symbol, or following the source);
- the inconsistent translation of other terms such as “click”, “log”, “Navigator” and “Support”, including the capitalisation of those words.

Finding a way of handling inconsistencies was actually a tough decision I had to make when calculating the error scores. Upon careful consideration, I decided to flag all occurrences as inconsistent whenever a term had not been translated consistently throughout the document. This decision was made in order to avoid introducing biases in the quality analysis that could favour specific types of translation suggestions, considering that the segments in the texts had suggestions of four different types. The failure to penalise a segment because the translation of a term had occurred first would mean favouring a suggestion type just because it occurred (by chance) first in the text. Moreover, the original IBM translation memory itself contained inconsistencies that were not removed when preparing the translation memories for the experiment. Therefore, the failure to penalise the segments where the most frequent form of a term translation occurred could unduly favour the segments that had benefitted from an originally better translation. For this reason, I considered that the best way to assess the decisions made by the translator based on the experiment conditions was to give one error point to all the segments involved in an inconsistency “case”. This is still not ideal, as inconsistencies were not evenly distributed among the different suggestion types, but the alternative solution of just not flagging inconsistencies would counter the quality assurance practices in the company, where inconsistencies were determined to be an important quality issue. Had the conditions of the experiment been different, other approaches would have been possible, such as the decision to consider the first-occurring translation of a term as consistent and to flag any discrepant translations of the same term that occurred later as inconsistent, as in Moorkens (2012: 76). Similarly, any repeated errors that were not inconsistencies were also penalised in every occurrence.

4.8.3. Perception

Translators' perceptions were measured using the interview methods described in section 4.6.4, namely semi-structured dialogues and retrospection with replay. The interviews were recorded, then transcribed and coded (Gorden 1992). Although some research tools such as NVivo[17] and ATLAS.ti[18] could help analyse qualitative data, I decided to process the interview data manually, as the number of interviews and the pieces of information to be looked for in each of them was relatively low. In the interview transcriptions, the passages where the translator was talking about each of the tasks or their attitude towards MT were highlighted according to a colour code. In order to better visualise the results, tables were created for each participant, where the verbal data was organised according to the three tasks (Scratch, Visual, Blind) and the three main variables: time (verbalised as "faster"), effort (verbalised as "more comfortable") and errors (verbalised as "better"). In some cases, as in the middle row in the example shown in Table 9, when one comment referred to both the Visual and the Blind tasks, it was inserted under the two corresponding columns.

Table 9. Coding of interview answers according to each task and variable (example)

Task Question	Scratch	Visual	Blind
Faster	3	2	1
Why?	Tienes que pensar más	No hay tanta memoria, se hacen rapiditos	Ya estaba hecho, era revisar
Better	[no mention]	1	1
Why?		Porque lo revisas todo igualmente. Deberían salir bien los dos.	
More comfortable	[no mention]	1	2
Why?		Porque sabes lo que es.	Como no sabes, tienes que ir viendo si…

[17] http://www.qsrinternational.com/products_nvivo.aspx

[18] http://www.atlasti.com

4.9. Ethical considerations

In the first stage of my research placement at MSS, I interviewed several people at the company in order to investigate the possibilities for conducting my experiments, such as the translator profiles, translation tools and text types that were available. I was also interested in gathering information on some of their business practices, in order to enhance my complementary skills, as set out in the guidelines of the TIME project. As a standard ethical practice, I asked my interviewees to sign a release form, which ensured anonymity of the data collected and granted me permission to use the information for research purposes. The release form for the interviews can be found in Appendix 10.

The main experiment involved keystroke and mouse logging, screen recording, voice recording, face recording and eye tracking. Participants were briefed on the type of data each of the methods was able to collect and on the main goals of the experiment. They were then asked to sign the release form that is presented in Appendix 9, granting their informed consent.

Since 2013, the Department of English and German Studies at the Rovira i Virgili University has had a regulation in place that requires researchers in the department to obtain approval from a commission prior to conducting any experiments involving human participants. At the time when my experiments were carried out, neither the university nor the department had any official requirements in this regard. The initiative to use the release forms stemmed from an established practice in the Intercultural Studies Group.

Chapter 5. Results

5.1. General quantitative results

In this section, results will be presented for the entire texts, as the idea is to compare the four tasks, and for the two preliminary tasks – the Copy and the Scratch tasks – it does not make sense to break up the texts in terms of suggestion types, because they contained no translation suggestions. For this reason, in what follows, translation time corresponds to the total time spent producing the translations (or the copy), including all the time spent for self-revising and proofreading the output. Similarly, typing effort and error score are also considered for the tasks as a whole. In section 5.4, our focus will move to the two main tasks and to the individual segments within the texts.

Table 10 shows the measured results for all ten participants and all four tasks when the entire texts are considered. *Translation time* is indicated as seconds per 100 source words. *Typing effort* is a percent ratio between the total number of relevant key presses and the total number of characters in the final target text. *Error score* is the total number of weighted errors per 100 source words.

Table 10. Translation time (seconds / 100 words), typing effort (%) and error score (weighted errors / 100 words) per participant in the four tasks

Participant	Translation time				Typing effort				Error score			
	Copy	Scratch	Visual	Blind	Copy	Scratch	Visual	Blind	Copy	Scratch	Visual	Blind
P01	146	257	191	200	110	102	14.0	11.9	0.0	3.8	1.0	1.0
P02	142	235	167	229	102	97.5	22.8	15.7	0.0	1.3	1.9	1.4
P03	151	324	215	193	107	103	50.0	13.3	0.0	5.5	4.3	4.5
P04	176	566	223	266	124	103	16.8	15.4	1.9	3.8	3.1	3.6
P05	157	259	121	157	114	106	12.4	11.5	1.9	5.1	4.3	4.2
P06	158	296	143	162	116	102	12.0	12.1	5.8	4.2	4.8	5.0
P07	158	613	232	334	105	109	13.0	14.5	1.9	3.8	3.3	3.1
P08	147	777	497	343	105	132	29.8	18.6	1.9	3.0	1.2	1.9
P09	175	344	139	139	103	153	22.6	6.4	0.0	8.9	5.4	5.0
P10	130	240	139	120	107	108	16.1	9.3	1.9	3.0	4.3	5.2

5.1.1. Translation time

The results for translation time in Table 10 are presented as a bar graph in Figure 7. They indicate that six out of the ten translators (P01, P02, P03, P04, P07 and P08) spent less time per word when carrying out the Copy task than any of the other tasks. P05 and P06 spent as much time on the Blind task as on the Copy task and they spent the least time on the Visual task. P09 spent more time on the Copy task than on the Visual or Blind tasks. P10 spent less time on the Blind task than on the Copy task.

Figure 7. Translation time (seconds / 100 words) for all participants in the four tasks

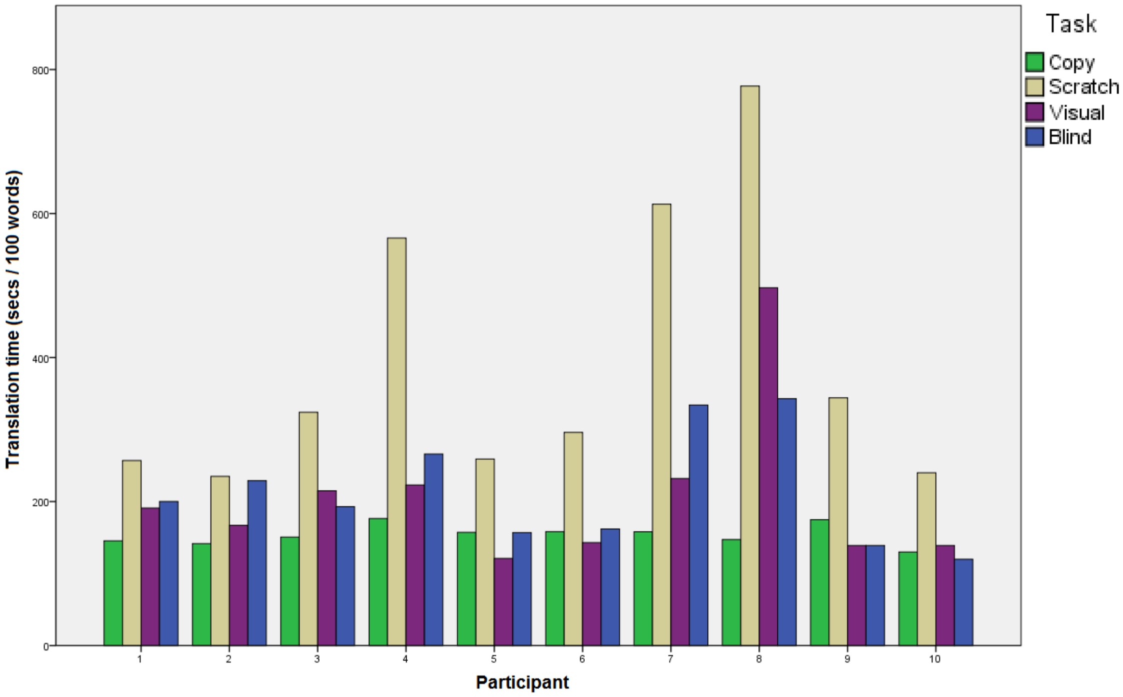
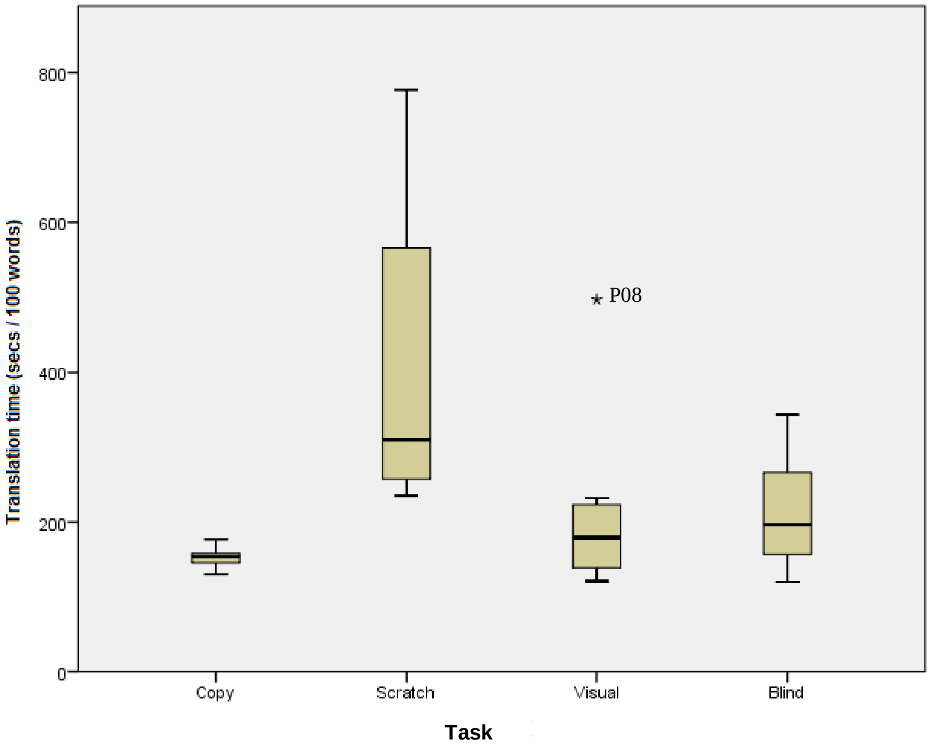


Figure 8. Dispersion of time data for the ten participants in the four tasks



The fact that the Copy task could take longer than some of the translation tasks might seem unexpected. However, it can be explained by the time-saving effect of the translation suggestions in the Visual and Blind tasks. Indeed, all translators spent the most time when translating from Scratch, which was the only translation task where no suggestions were provided. The only exception is P02, who spent almost as much time on the Blind task as on the Scratch task.

Now comparing the times between the Visual and the Blind tasks, five participants spent more time on the Blind task (P02, P04, P05, P06 and P07), three participants spent more time on the Visual task (P03, P08 and P10) and two participants had differences of below 5% between those tasks (P01 and P09).

The data for translation times in Table 10 are also displayed as box plots in Figure 8, to show the dispersion of data in the sample. It is interesting to notice that although most participants have very similar typing speeds – indicated by the box plot for the Copy task, with minimal variance – their speeds when translating from Scratch vary enormously. When it comes to the two translation tasks that include translation suggestions, the Visual and the Blind tasks, their speeds come closer together again. P08 stands out as an outlier in the Visual task (indicated by the asterisk in the box plot) and is also responsible for the maximum values in the Scratch and Blind tasks. This translator will be analysed later (see sections 6.2.5 and 7.4.1).

5.1.2. Typing effort

As explained in Section 4.3.2, typing effort is calculated as the number of keystrokes used to produce a particular segment divided by the total number of characters in the resulting segment. The results for typing effort presented in Table 10 are shown as a bar graph in Figure 9. They indicate that all ten translators typed at least 100 percent of the characters required to produce the target text in the Copy and Scratch tasks, except for P02, who typed slightly less than 100 percent in the Scratch task. As a reminder, in the Copy task the start text was provided only in print form, whereas for the Scratch task the translators had the source text both in print form and in the translation tool. Each target segment was populated by default with the source text in the Scratch task, which allowed six translators to type less in the Scratch task than in the Copy task, as they could use words that were identical in English and Spanish, such as product names. It also explains how P02 even managed to type less than 100 percent of the characters required to produce the translation from scratch. But the presence of source text in the target segments also accounts for the opposite phenomenon for P09, who had a much higher percentage of typing effort in the Scratch task. This participant used the Delete key to delete the remnants of the source text in the editing area, instead of deleting blocks of texts with key combinations. We should remember that each press of the Delete key also counts as one keystroke in the calculation of typing effort. P07 and P08 (and P10, to a lesser degree) also typed more in the Scratch task than in the Blind task, but in their case this happened because they changed their minds on multiple occasions and decided to modify parts of the translations they had already produced (deleting plus rewriting).

In the Visual and Blind tasks, all translators had a much lower typing effort than in the Copy and Scratch tasks. Most of the translators typed more in the Visual task than in the Blind task, except for P06, who typed virtually the same percentage in both tasks, and P07, who actually typed less in the Visual task.

The box plot in Figure 10, also based on the data in Table 10, shows the dispersion of typing data in my sample. Despite some outliers in the Scratch and Visual tasks, there is much less dispersion in the data for typing effort than for translation time. This is an initial indicator of a lack of correlation between these two variables among the ten participants, as will be shown in section 5.1.5.

5.1.2.1. Typing effort, P09 normalised

It was mentioned above that P09 used the Delete key to erase the source text in the Scratch task, which made this participant the one who typed the most in that task. When she deleted her own text (fixing typos) she always used the Backspace key, so it is safe to just eliminate all Delete keystrokes in order to normalise her data. The result is a drop in the figures for her typing effort from 153% to 93% (Figure 11). After doing this, she is no longer an outlier in the boxplots (Figure 12).

Figure 9. Typing effort (%) for all participants in the four tasks

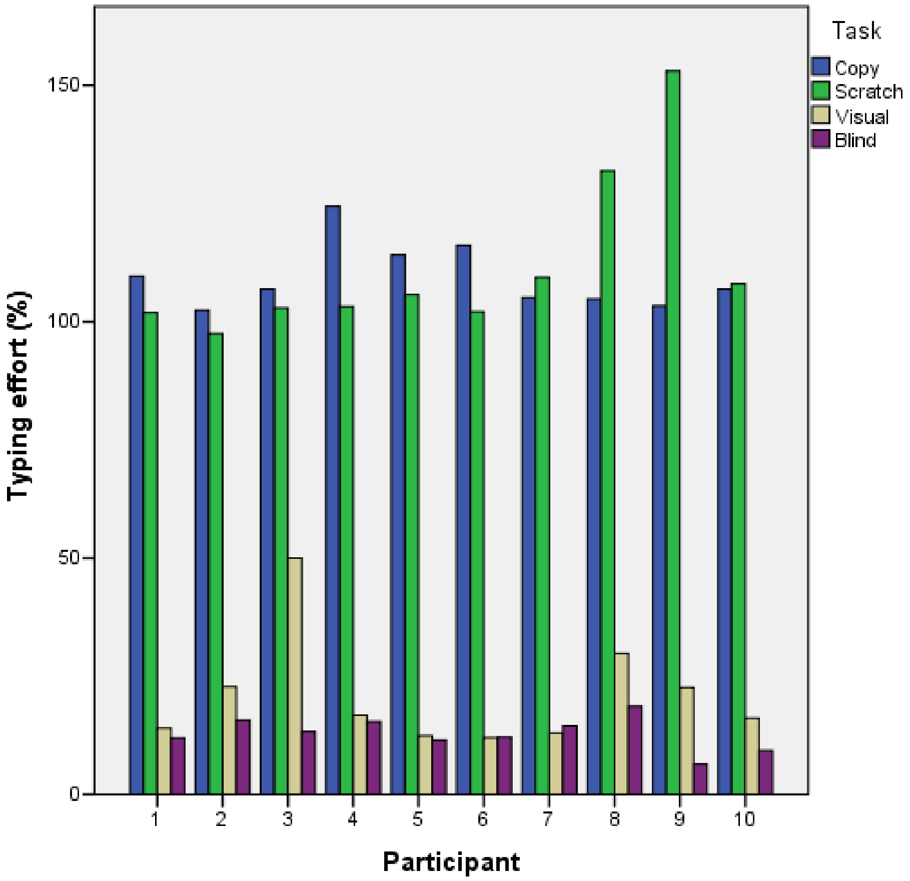


Figure 10. Dispersion of typing effort for the ten participants in the four tasks

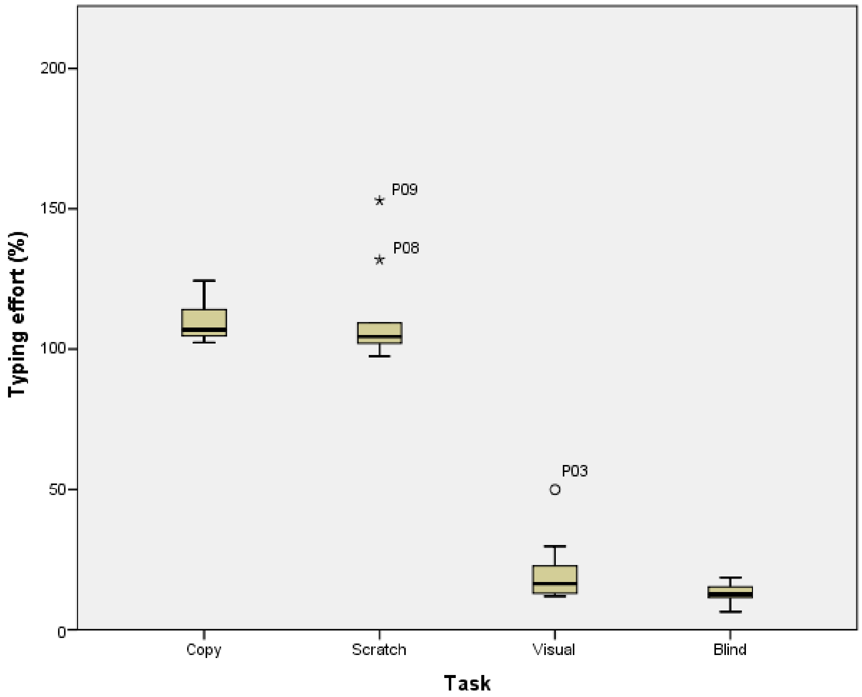


Figure 11. Typing effort (%) for all participants in the four tasks, with normalised data for P09

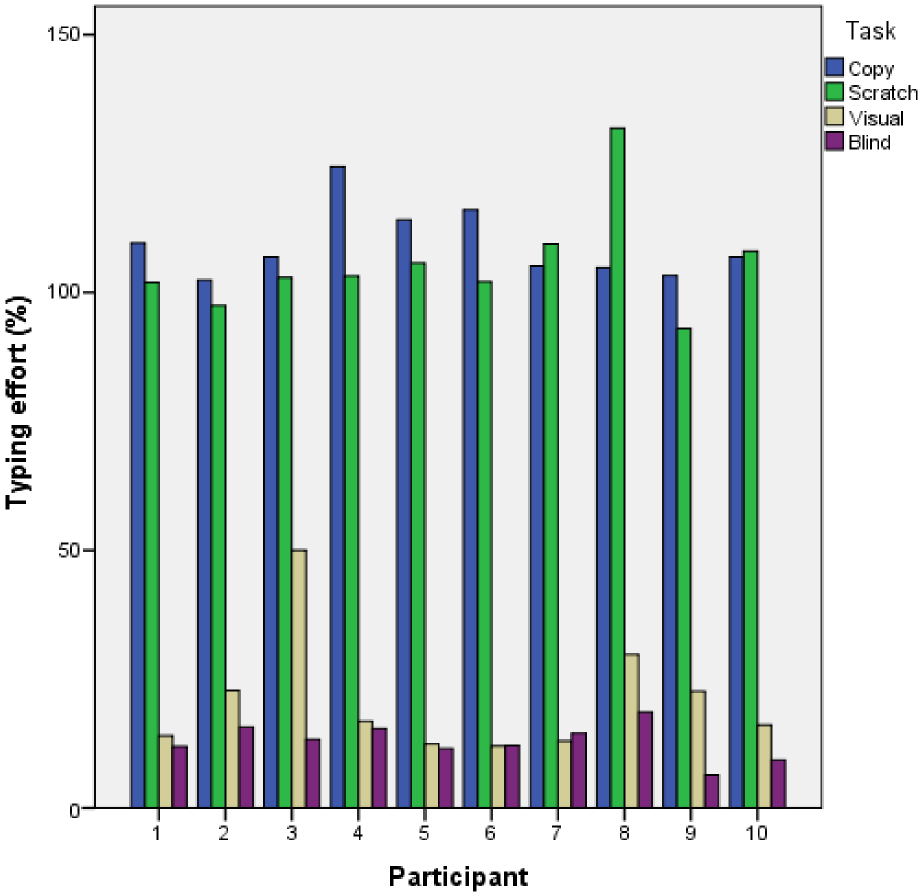
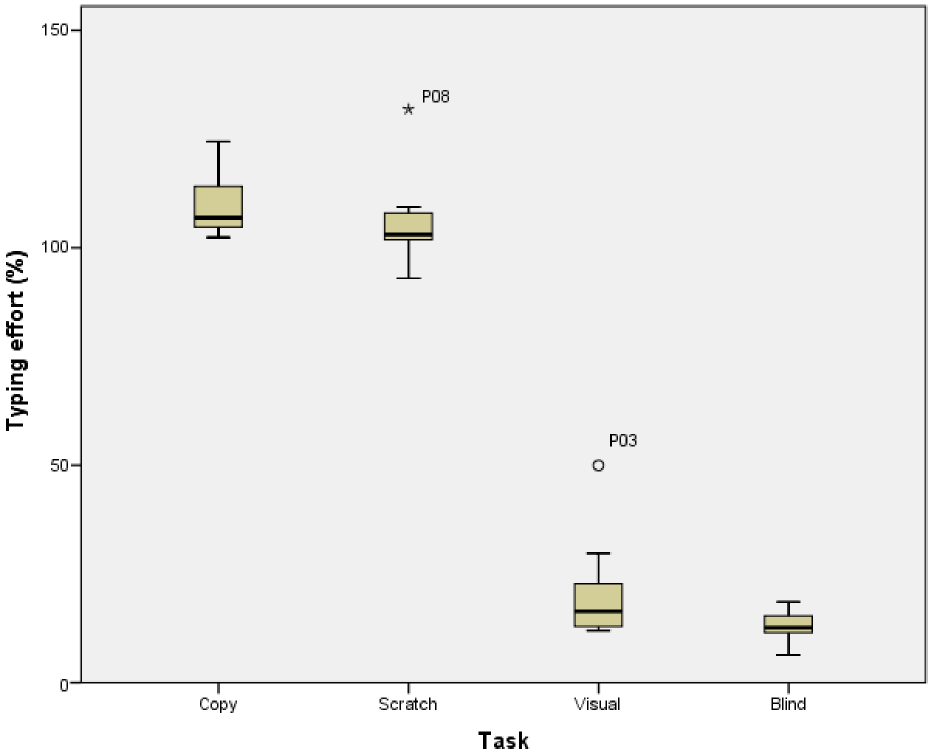


Figure 12. Dispersion of typing effort for the ten participants in the four tasks, with normalised data for P09



5.1.3. Error score

The bar graph in Figure 13, again based on the data presented in Table 10, shows the error scores for the ten translators in the four tasks. In the Copy task, four translators (P01, P02, P03 and P09) made no errors, five translators (P04, P05, P07, P08 and P10) made 1.9 errors / 100 words (they actually made one error in the 52-word file) and one translator (P06) made 5.8 errors / 100 words (he made one light error and one severe error, weighted as two, in the entire text). Except for P06 and P08, the Copy task was the task in which translators made the fewest errors.

The relative number of errors between the other three tasks varied according to the participants, although seven of them made more errors in the Scratch task than in the Visual or the Blind tasks. From the box plot in Figure 14, we can see that the median for the three tasks is virtually the same, at around 3.8 errors per 100 words. The mean, however, is higher for the Scratch task, at 4.2 errors per 100 words, due to the high number of errors produced by P09. As a general observation, the presence of translation suggestions in the Visual and in the Blind tasks did not have an impact on the number of errors as compared to the Scratch task. The same is true for the presence of translation metadata in the Visual task, which did not account for a reduced number of errors in this environment, when considering the texts as a whole. In sections 5.4.6 and 5.4.7, we will see that the error scores are only affected by the type of translation suggestion, not by the presence or absence of metadata (task type).

Figure 13. Error scores (errors / 100 words) for all participants in the four tasks

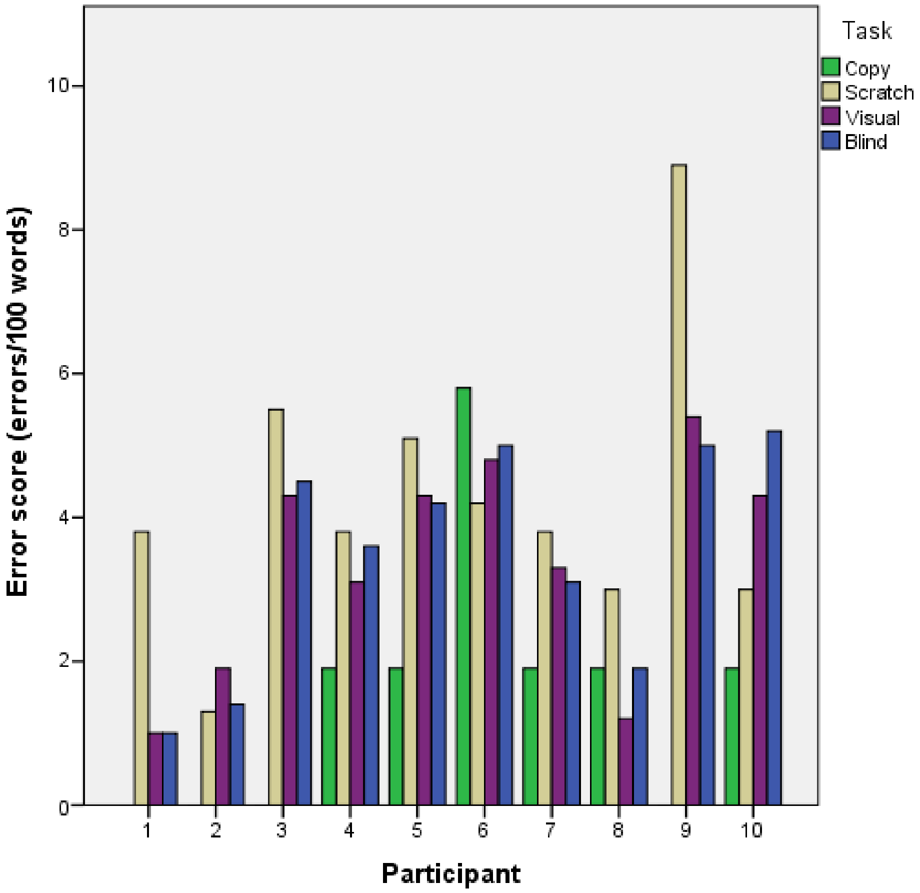
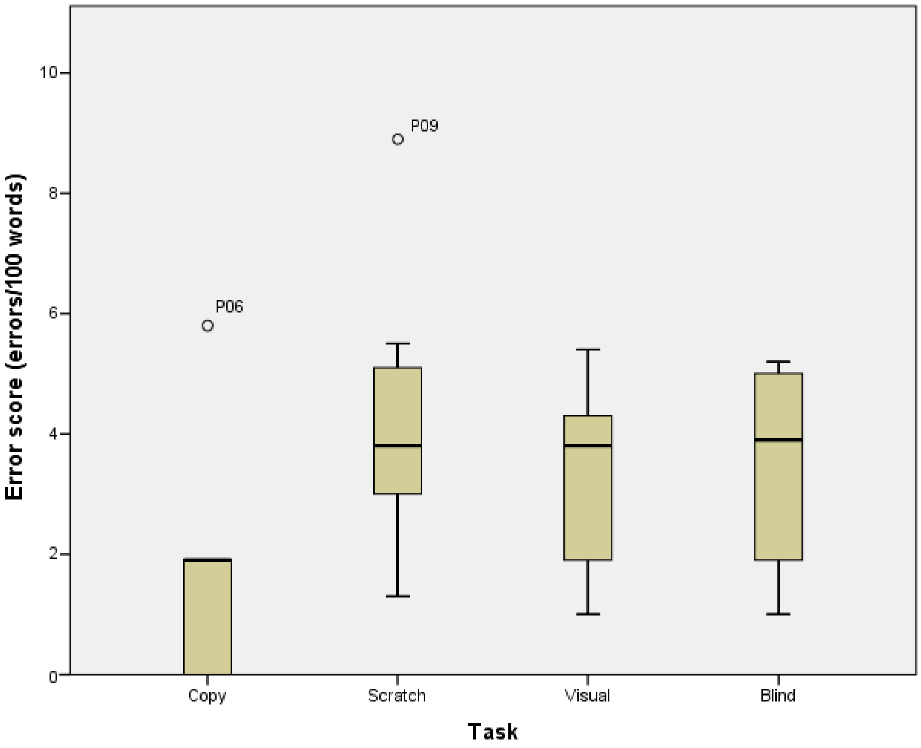


Figure 14. Dispersion of error scores for the ten participants in the four tasks



5.1.4. Correlation between tasks

I have mentioned that one of the goals of the Copy task was to measure the participants' baseline performance, i.e. to measure how much time they spent, how much typing effort they invested and how many errors they produced in a non-translation task as compared to the three translation tasks.

In order to test for potential statistical correlations, I first checked the data distributions for normality. According to the Shapiro-Wilk test, the data for translation times in the Scratch and Visual tasks, for typing effort in the Scratch and Visual tasks, and for error scores in the Copy task can be considered non-normally distributed, as indicated by the significances below 0.05 in Table 11. The data for the remaining variable–task combinations are assumed to be normally distributed.

Table 11. Results of the Shapiro-Wilk test of normality

Variable	Task	Statistic	df	p
Translation time	Copy	0.953	10	0.708
	Scratch	*0.800*	*10*	*0.014*
	Visual	*0.698*	*10*	*0.001*
	Blind	0.912	10	0.297
Typing effort	Copy	0.872	10	0.104
	Scratch	*0.716*	*10*	*0.001*
	Visual	*0.765*	*10*	*0.005*
	Blind	0.982	10	0.977
Error score	*Copy*	*0.750*	*10*	*0.004*
	Scratch	0.893	10	0.182
	Visual	0.913	10	0.302
	Blind	0.898	10	0.210

Note: Rows in italics indicate non-normally distributed data ($p \leq 0.05$).

My initial approach for investigating potential correlations was to run parametric tests (Pearson) for the normally distributed data and non-parametric tests (Kendall and Spearman) for the non-normally distributed data. Since the qualitative differences in the results were minor between the different tests, only the results for Spearman's tests are presented here. At $\alpha = 0.05$, no statistically significant correlation was found between any of the variables in the Copy task and the corresponding variables in any of the three other tasks, as can be seen in Table 12.

Table 12. Spearman correlation tests between the Copy task and the three translation tasks

Variables	Tasks	Copy vs. Scratch	Copy vs. Visual	Copy vs. Blind
Time	Coefficient	0.612	0.030	0.079
	p (2-tailed)	0.060	0.934	0.829
Effort	Coefficient	-0.28	-0.559	-0.152
	p (2-tailed)	0.434	0.093	0.675
Errors	Coefficient	-0.122	0.262	0.281
	p (2-tailed)	0.736	0.464	0.431

These results show that, for each translator, the time spent translating (in any of the three translation tasks) had no direct correlation with the time spent copying, the percentage of edits while translating had no direct correlation with the edits made while copying, and the number of errors made while translating had no direct correlation with the errors made while copying. Albeit counter-intuitive, this is in accordance with the results of other studies. In their research comparing touch typists with non-touch typists, Sharmin et al. (2008) found that, although the difference in typing style had a significant effect on eye fixations on the screen, it did not correlate significantly with how the participants responded to time pressure and text complexity. Their results seem to indicate that being a faster typist does not imply being a significantly faster translator. In section 6.2.9, I argue that this is due to the time spent on other tasks during the entire translation process, of which typing constitutes a small part.

Significant correlations could be found, however, between some of the other tasks (see Table 13). Spearman's tests indicate a significant correlation for translation time between the Visual and the Blind tasks ($\rho = 0.903$, $p < 0.001$) and significant correlations for error scores between the Scratch and the Visual tasks ($\rho = 0.633$, $p = 0.05$) and between the Visual and the Blind tasks ($\rho = 0.906$, $p < 0.001$). No correlation was found for typing effort (edits) between any of the tasks.

Table 13. Spearman's correlation tests between the three translation tasks

			Time			Edit			Errors		
			Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
Time	Scratch	Coefficient		0.624	0.552						
		p (2-tailed)		0.054	0.098						
	Visual	Coefficient			*0.903*						
		p (2-tailed)			*0.000*						
	Blind	Coefficient									
		p (2-tailed)									
Edits	Scratch	Coefficient					0.103	-0.224			
		p (2-tailed)					0.777	0.533			
	Visual	Coefficient						0.394			
		p (2-tailed)						0.260			
	Blind	Coefficient									
		p (2-tailed)									
Errors	Scratch	Coefficient								*0.633*	0.480
		p (2-tailed)								*0.050*	0.160
	Visual	Coefficient									*0.906*
		p (2-tailed)									*0.000*
	Blind	Coefficient									
		p (2-tailed)									

Note: Cells in italics indicate significant results ($p \leq 0.05$).

5.1.5. Correlation within tasks

Now I take each of the tasks individually and look for potential correlations between the variables among the ten translators. The results of Spearman's correlation tests are as presented in Table 14. They indicate a positive correlation between translation time and typing effort in the Scratch task ($\rho = 0.648$, $p < 0.05$) and in the Blind task ($\rho = 0.903$, $p < 0.001$), suggesting that the more one typed the longer it took to complete the task, as intuitively expected. However, no correlation between translation time and typing effort was found in the Visual task, suggesting some effect associated with the presence of metadata in this task. The results also show a significant negative correlation between translation times and error scores in the Visual task ($\rho = - 0.687$, $p < 0.05$) and in the Blind task ($\rho = - 0.770$, $p < 0.05$), now suggesting that the more time one spent on the task, the fewer errors one made, as might also be expected. Finally, a significant negative correlation was found between typing effort and error scores in the Blind task ($\rho = - 0.648$, $p < 0.05$), suggesting that the more one typed the fewer errors one made, although this was not the case in the other tasks.

Table 14. Spearman's correlation tests for the three variables within each task

Tasks	Variables	Time vs. Effort	Time vs. Error	Effort vs. Error
Copy	Coefficient	0.383	0.281	0.534
	p (2-tailed)	0.275	0.431	0.112
Scratch	Coefficient	*0.648*	0.240	0.228
	p (2-tailed)	*0.043*	0.504	0.527
Visual	Coefficient	0.382	*-0.687*	-0.316
	p (2-tailed)	0.276	*0.028*	0.374
Blind	Coefficient	*0.903*	*-0.770*	*-0.648*
	p (2-tailed)	*0.000*	*0.009*	*0.043*

Note: Cells in italics indicate significant results ($p \leq 0.05$).

5.2. General interview data

After completing the four tasks, each translator was interviewed, as explained in section 4.6.4. The following three main questions were asked in relation to the three translation tasks (Scratch, Visual, Blind):

1) Do you think you translated *faster* in any of the environments? If so, in which one?
2) Do you think the quality of your final translation was *better* in any of them? If so, in which one?
3) In which environment did you feel more *comfortable* working?

The resulting dialogues lasted 7 minutes on average, with a minimum of 2 minutes (P10) and a maximum of 32 minutes (P01). The retrospections lasted between 11 minutes and 18 minutes, with an average of 14 minutes, but contained long periods of silence. Both kinds of interviews were transcribed, resulting in 3 hours and 4 minutes of recordings and a total of approximately 10,500 words by the participants. For two participants it was not possible to carry out the retrospection, either because of technical reasons (P05) or because the participant refused to do it (P08).

More detailed information about the interview data will be given in section 5.3. The focus of the current section is to check for correlations between the measured results presented in section 5.1 and the perceived results obtained from the interviews. In order to make the qualitative and quantitative data comparable, the approach used was to rank each variable in each of the tasks for each subject, both as measured and as perceived, and then to compare the rankings.

As a result of coding the interview data as explained in section 4.8.3, Table 15 shows how the translators perceived their performance after the translation tasks. “1” indicates the lowest rank of the variable as perceived by the translator, e.g. Time = “1” means that the corresponding participant mentioned this was the fastest task (lowest time). The blank cells in the table represent data for which no clear answer was given in the interview. As a general observation, the table shows that all participants thought they made fewer errors and invested less effort in the Visual task than in any of the two other translation tasks (or at most the same amount of errors and the same level of effort), and that most of them considered they spent the least time on the Visual task.

Table 15. Perceived times, effort and errors as ranks in the three translation tasks (interview data)

Participant	TIME			“EFFORT”			ERRORS		
	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	3	3	1		1		1	1	1
P02	3	1	1		1	2		1	1
P03		1		3	1	1	3	1	1
P04		1			1			1	3
P05		1			1			1	
P06	2	1	3	2	1	3	2	1	3
P07		1		2	1	3	2	1	3
P08	3	2	1		1	1		1	2
P09		1		2	1	3			3
P10		1	2		1			1	1

In the following sections we will compare this perceived data with the measured for each of the dependent variables.

5.2.1. Comparison between quantitative and qualitative data

In Table 16, the quantitative results shown in Table 10 are also converted into ranks. The Copy task is not taken into account, as in the interviews the translators were only asked to compare the three translation tasks. For each particular translator and for each variable in Table 10, the task with the lowest number is assigned rank 1 in Table 16, the task with the highest number is assigned rank 3 and the intermediary task is assigned rank 2. When the difference between two tasks is not relevant, considering a deviation of ± 5 percent, the same rank is assigned to more than one task, giving preference to the extreme ranks 1 and 3. The reason for preferring the extremes is that this corresponds better to the types of answers available from the interviews (e.g. the “fastest” task vs. the “slowest” task), and it possibly also reflects better the way human perception works.

Table 16. Measured times, edits and errors as ranks in the three translation tasks (process data)

Participant	Translation time			Typing effort			Error score		
	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	3	1	1	3	2	1	3	1	1
P02	3	1	3	3	2	1	1	3	1
P03	3	2	1	3	2	1	3	1	1
P04	3	1	2	3	2	1	3	1	3
P05	3	1	2	3	1	1	3	1	1
P06	3	1	2	3	1	1	1	3	3
P07	3	1	2	3	1	2	3	1	1
P08	3	2	1	3	2	1	3	1	2
P09	3	1	1	3	2	1	3	1	1
P10	3	2	1	3	2	1	1	2	3

Table 16 indicates that all translators spent the most time and made the most edits (represented by the number 3) when translating from Scratch. The same cannot be said about the errors, since three of the translators made the fewest errors when translating from Scratch. The table also shows that most translators performed the fewest edits in the Blind task (or as few edits in the Blind task as in the Visual task), except for translator P07, who typed less in the Visual task.

In the following sections, the results in this table will be compared with the translators' perception.

5.2.1.1. Time

The time measured per 100 words was consistently higher when translating from scratch for all ten participants (Table 16). This is generally in accordance with their perception (Table 15) – despite some missing data – except for one translator (P06), who thought he spent less time translating from scratch than he did on the Blind task. For the seven translators who thought they were faster in the Visual task than in the Blind task (P03, P04, P05, P06, P07, P09, P10), all but two (P03, P10) were indeed faster. For the two translators who thought they were faster in the Blind task than in the Visual task (P01, P08), P08's perception corresponded to his measured times, whereas the difference in time for P01 was not noticeable between the two tasks. The only participant who thought he was as fast in the Visual as in the Blind task (P02) was actually much faster in the Visual task.

5.2.1.2. Comfort vs. effort

As indicated in Table 16, the Blind task was the condition in which the translators typed the least, except for one translator (P07), who typed less in the Visual task. Two translators (P05, P06) typed as much in the Blind task as in the Visual task. A simple comparison of the middle columns in Table 16 and Table 15 reveals no coincidence between the measured edits and the perceived "effort" while performing the task. This could be attributed to any of the factors mentioned in Section 5.3, but in this case, the discrepancies in the results are probably due to a poorly formulated question. The quantitative variable being measured as an indication of effort was the amount of editing, which is a simple measurement of physical effort, while in the interviews the translators were asked about the task in which they felt more "comfortable". It turns out that typing effort and the feeling of "comfort" while performing a task are not directly comparable, as had been previously suggested by Koponen et al. (2012: 20): "keystrokes, while very useful as a way to understand how translators work, may not be an appropriate measure to estimate cognitive effort".

5.2.1.3. Errors

Table 16 shows that 70 percent of the translators made the most errors when translating from scratch, which might indicate their reliance on translation suggestions, after many years of practice working with translation memories. There was no clear difference between the Visual and the Blind tasks in terms of error rates, although most translators thought they made the fewest errors in the Visual task, or as few errors in this task as in the Blind task. P09 did not distinguish explicitly between the Visual task and translating from scratch. Their perceptions corresponded to the reviewers' quality assessment in 70 percent of the cases, whereas two translators (P02, P06) actually made the most errors in the Visual task and one translator (P10) made more errors in the Visual task than when translating from scratch.

In general terms, the participants tended to overrate the Visual task (with translation suggestions and metadata), as the measured data shows that their performances on this task was largely comparable to their performances on the Blind task (with translation suggestions and no metadata).

5.3. Additional information from the interviews

A major goal of the interviews was to let participants express their priorities. This was achieved through a relatively free dialogue format, which was responsible for some of the missing data in Table 5, but also allowed other factors to come into play that had not been included as the main variables in the study.

5.3.1. Translation vs. revision vs. post-editing

The interviews indicate a clear difference in the way translators perceived the two main translation tasks. All participants except one made a clear distinction between “translate”, for the Visual task, and “revise” or “proofread” (“revisar” in Spanish) or “post-edit”, for the Blind task. The quantitative data support this perception, as they show many more iterations per segment in the Visual environment, as if the translators were first translating, then self-revising. In the Blind environment, which they considered to be revising or post-editing, they completed the task in a single round. This difference made seven of the translators feel that they had performed a regular revision (on text that had been translated or proofread by another translator) when working in the Blind task (my translations here and throughout):

> P10: I’m very much used to working the first way, to translate. I had never done the other task before actually, to find everything at 100% and to revise it.
>
> P02: The other one was already done, we just had to revise.
>
> P01: Post-editing, a revision that had already been done and that I had to revise.

For these participants, the text they were “revising” was in principle better than the text they had in the Visual task:

> P08: We assume that in theory it should be better.
>
> P04: There was a lot of [translation] memory and it was quite good compared with other folders.
>
> P07: We could notice some segments had been leveraged from the memory... they were better, I didn’t have to change much.

Only one participant (P05) felt she was “translating” when performing the Blind task: she actually talked about both tasks in terms of the presence or absence of metadata on the translation suggestions.

5.3.2. The role of translation suggestions

Seven translators acknowledged the usefulness of translation suggestions (as opposed to translating from scratch):

> P02: Because [when you translate from scratch] you have to think more.
>
> P03: It always helps to have pre-translated stuff or when there is something previous that is useful, because if you translate everything from scratch, you always make mistakes, [it's a little] more difficult. Having something as a basis is always welcome.
>
> P06: When you have a suggestion from the memory, you insert it and if you change a word, maybe you go faster too, with some memory. [Pause] Translating 500 words with memory suggestions is faster than from scratch…
>
> P07: Because you have an external aid from previous memories and machine translation [...] you always go faster. [...] it is always better to have some help.
>
> P08: When there is a suggestion, you go much faster.
>
> P09: [The Visual task] had many fuzzies at 95%, 85%, so it is very easy to detect where the small changes are, and it is very useful.

One of those participants (P09), however, pondered that it might be easier to translate from scratch:

> P09: It is easier to translate from scratch, because I don't have to look at anything. And I don't need to check if what is suggested is correct or not, or if it's in the right order or in the wrong order.

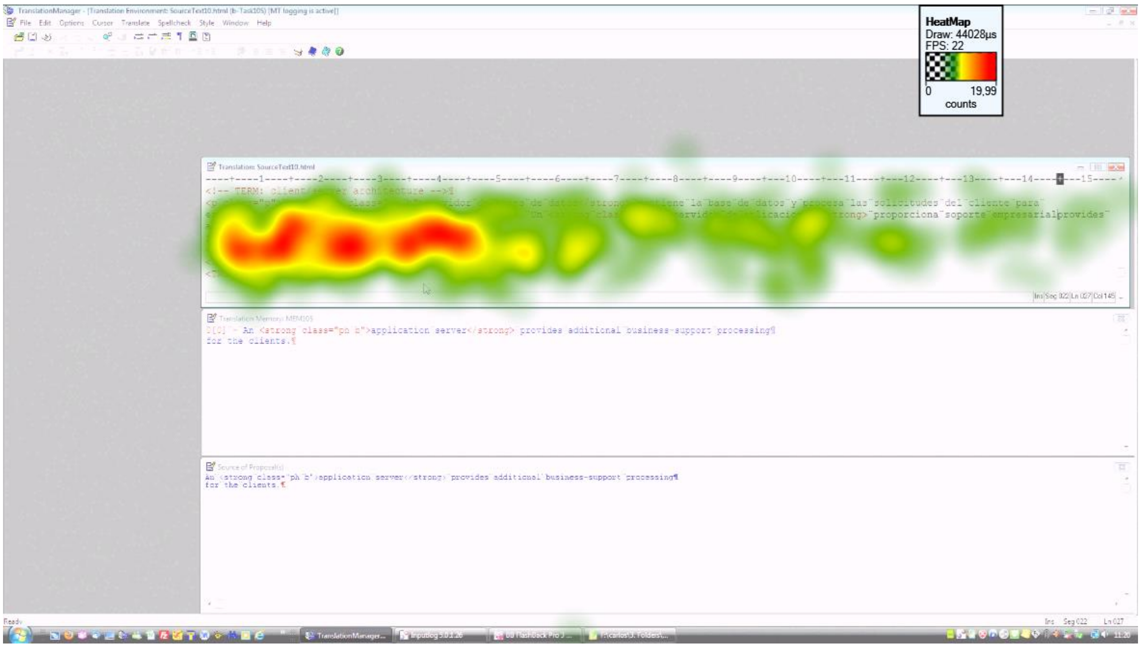
Along the same lines, P01 said:

> P01: I don't think it's especially faster having the memory, because when you translate from scratch, one advantage I can see is the vocabulary, but the other is that there is no suggestion to look at, no differences to check for between one sentence and the other. [...] I think I compensate what I use – the help from the memory – with the time I spend checking the passage, checking for differences.

The eye-tracking data indicate that P09 and P01 indeed fixated on a single area in the screen while translating from scratch, as there are no translation suggestions or translation metadata to look at. P01's heatmap in Figure 15 below shows her fixations only on the part of the tool where the source and target texts are (the segment initially contains the source text, which is gradually replaced by the translators as they type). This particular subject did not look at the keyboard at all.

P09's heatmap is very similar to P01's. The main differences are that she looks up a term in the glossary pane and then self-revises at the end, as can be seen by watching her screen recording. More examples of eye-tracking data analysis are presented in section 5.5 and will illustrate the differences in gaze behaviour between the Visual and Blind tasks.

Figure 15. Heatmap of P01 while translating from scratch



Translation: SourceText10.html

<!-- TERM: client server architecture -->¶

Translation Memory: MEM05

[C] - An <strong class="ph b">application server</strong> provides additional business-support processing¶
for the clients.¶

Source of Proposal(s)

An <strong class="ph b">application server</strong> provides additional business-support processing¶
for the clients.¶

The three remaining participants did not make any specific comparisons involving the translation from scratch.

5.3.3. The role of metadata

Even if the translators did not consider the metadata to be the main distinction between the Visual and the Blind tasks in their comments, they demonstrated awareness of how translation metadata could help them:

> P01: If I see a fuzzy match, the first thing I'll look at is the Source of Proposal. For me it's easier with a memory, with fuzzy matches, with information on whether it comes from MT or from fuzzy or whatever, because it allows me to look at it in one way or another.

> P02: If you see that it's 100%, that it's not machine translation, then, in principle, in an everyday translation, when you go fast, you don't even look at it. You assume it's correct or that you translated it yourself before. [...] A fuzzy match, if I see that everything is translated and there is only one word that changes, I change that word, I don't even look at the rest.

> P04: Because you can't see below where it comes from... [when there is no metadata]
>
> P05: TM/2 indicates the fuzzy matches... it highlights what is missing, what is extra, what has changed.
>
> P06: You always look at what has changed and you change there. [...] You didn't even need to read the sentence, you just had to change a word that was highlighted and that's it.
>
> P08: When it's pre-translated you don't have... you don't know the quality of the suggestion; in contrast, when you have the memory, you know if it's an MT suggestion or if it comes from a... from another publication. TM/2 indicates if it's an Exact Match or if it's an MT suggestion or if it's a fuzzy match... [...] so if you know it's MT, you look at it with more... respect. Conversely, if you know it's a fuzzy match, since you know it has been checked by a human translator, it gives you more confidence. Sometimes you just look at what has changed. On the other hand, when you have it pre-translated, I don't know where it comes from... I would prefer to know... the environment where you see the suggestion, if it's machine translation, if it's... or if it comes from another publication that has been checked by somebody else. I think it's better to have the information, because it tells you what has changed; so if you know what's changed, you focus more on what's changed. Your natural tendency is to trust more what appears as unchanged.
>
> P09: The second one [Visual] had several fuzzies at 95%, 85%, so it's very easy to detect where the small changes are, and it's very useful. [...] If you look at the suggestion, since it tells you exactly what the changes are, it's easier to detect. [...] For me it's much easier to upload or to edit.

Morado Vázquez (2012: ii–iii) obtained similar feedback from the translators in her study: "In terms of participants' attitude towards the metadata received, most of the participants did not find it distracting, and the majority of them would prefer a translation memory which contained metadata." It is worth noting, however, that one translator in my experiment (P08) stated, "the environment that gives you more information is, at the same time, more complex".

One participant (P10) mentioned that he missed another type of meta-information that was absent in the Blind environment, according to his tool settings:

> P10: It's very clear for me that the blue segments are the ones I have to fix, the ones I have to work on, and everything was black here.

Three translators highlighted the importance of reference sources built into the translation tool, namely the client glossaries with product-specific terminology. P07 did not mention metadata and focused her comments on the opposition between translation memory and machine translation.

5.3.4. The perception of machine translation

In general, the participants had mixed feelings about machine translation. Although in some cases they criticised it as being poor, they also recognised that some machine-translated segments were "almost perfect" and that MT helped them increase productivity.

Two translators (P06, P09) felt the text in the Blind task contained more machine-translated segments than the text in the Visual task, although the translators were told that both texts actually had the same distribution of suggestion types, and only 25% of the suggestions were actually machine translation feeds (see Sections 4.3.5 and 4.5.6). Therefore, in their comments the translators made statements about the (presumably lower) quality of the translation suggestions based on their assumption that the suggestions come from machine translation:

> P06: In the revision task, since they come from machine, they are always faulty.
>
> P09: [The Blind task] is mostly machine, so it takes me longer to think about what changes [...]. I do have to keep thinking what the core of the segment is and to change it.

He et al. (2010) and Guerberof Arenas (2013: 87–88) also show evidence that translators tend to trust fuzzy matches more than they trust machine translations and that in many cases subjects are not able to tell TM suggestions from MT suggestions.

5.3.5. Task familiarity

Eight out of the 10 participants (P01, P02, P04, P05, P06, P07, P09, P10) reported being more comfortable tackling the Visual task, even when some believed the Blind task could be faster. The other two participants (P03, P08) were equally comfortable working in the Blind task. P08 found the Blind task "more simple":

> P08: You look at the English, the Spanish and that's it. [...] In the other one, you have to look at the English, the Spanish, and sometimes choose among five suggestions – not the case in this experiment though, where you had only one suggestion.

The main reason given by the translators (mentioned by 7 out of 10 translators) for feeling more comfortable and actually preferring the Visual task was that they were very "used to" or "more familiar with" (in Spanish: "acostumbrado a", "familiarizado con", "habituado a") the Visual task, while the Blind task was new to them. Another reason given by the translators (3 out of 10) for preferring the Visual task was that they felt more confident and secure in this environment. It is unclear in some statements whether this feeling of confidence is only related to task familiarity or also to the metadata or to any other characteristics present in the Visual task:

> P01: I prefer to translate with a memory. [...] For me it's more comfortable, it makes me feel more confident.

> P04: Surely because this is what I've been doing for IBM lately, [I feel] more confident, maybe more familiar with it.

> P08: If you know it's a fuzzy match, since you know it has been checked by a human translator, it gives you more confidence.

These findings are in agreement with other studies that have also identified *familiarity* as an important factor affecting the acceptance of MT and of translation technologies in general among translators (Webb 1998; Wallis 2006; Dillon and Fraser 2006; Lagoudaki 2008; Doherty and Moorkens 2013; Guerberof Arenas 2013).

5.3.6. Different strategies

Since all the participant translators were used to doing revisions in IBM TranslationManager, where the text to be revised comes pre-translated (but with metadata on the provenance of existing translations), their feeling of unfamiliarity or lack of confidence with the Blind task can probably be explained by the absence of metadata in this task. This suspicion is reinforced by several statements in which translators explain that they use different strategies for exact matches, fuzzy matches and machine translation:

> P01: If I see it's an "m" [machine translation], I read the sentences from A to Z, or I go and check for some things or I look for some things or for other things. If I see a fuzzy match, the first thing I'll look at is the Source of Proposal. For me it's easier with a memory, with fuzzy matches, with information on whether it comes from MT or from fuzzy or whatever, because it allows me to look at it in one way or the other. If I see a fuzzy match, I look at the Source of Proposal; if I see an MT, that is, if I see an "m", and it gives me the impression that the sentence is more or less correct, then I insert it and, depending on the case, I fix it, because sometimes the sentence is almost entirely perfect.

> P02: If you know it's... you look at it differently. If you see that it's 100%, that it's not machine translation, then, in principle, in an everyday translation, when you go fast, you don't even look at it.

These testimonials are in accordance with feedback provided by participants in other studies (O'Brien 2006a: 198), as different types of translation tasks seem to activate different translation strategies and to require different allocation of cognitive resources (Lörscher 1991; Jääskeläinen 1993; House 2000; Carl et al. 2010; Hvelplund 2011; Dragsted 2012). The fact of knowing which type of suggestion is being dealt with when processing a segment could reduce cognitive load and account for the reported feeling of comfort.

5.4. Quantitative results by suggestion type

In the previous sections in this chapter, we have looked at each of the tasks considering the texts as a whole. In this section, I will analyse the results of my experiment looking at the individual segments of each text, as those segments offered different types of translation suggestions and could affect translators' performances in different ways. The focus will move to the Blind and Visual tasks only, as the Copy and the Scratch tasks offered no translation suggestions.

As in the previous sections, my dependent variables are Translation Time, Typing Effort and Error Score (here capitalised with title case for the sake of clarity in the statistical analysis). These are scalar, numeric variables. The independent variables in this study are of two types: categorical variables (factors) and numeric variables (covariates). The two main independent variables (taken from my hypotheses) are categorical: Task, which has two levels (Visual or Blind), and Suggestion Type, which has four levels (Exact Match, High Fuzzy Match, Low Fuzzy Match and Machine Translation). The two levels of Task correspond to the presence (Visual) or absence (Blind) of translation metadata in the translation suggestions. In addition to the main independent variables, other variables contain information about the experiment design – Text, Task Order and Text Order –, or about the participants – Gender, Age, Experience –, as well as the baseline measurements for task time, typing effort and error score in the Copy task (here named Copy Time, Copy Effort and Copy Errors). Gender, Task Order and Text Order are categorical variables (factors), each with two possible levels. Age, Experience, Copy Time, Copy Effort and

Copy Errors are numeric in nature (covariates). Table 17 summarises the variables that are analysed in this and the following sections.

Table 17. Variables included in the statistical analysis

Role		Name	Type	Measurement / Levels
Dependent		Translation Time	numeric	seconds / 100 words
		Typing Effort	numeric	typed chars / target chars (%)
		Error Score	numeric	errors / 100 words

Independent	Primary	Task (metadata)	categorical	V = Visual (present),
				B = Blind (absent)

		Suggestion Type	categorical	E = Exact Match
				H = High Fuzzy Match
				L = Low Fuzzy Match
				M = Machine Translation

	Secondary	Gender	categorical	M = Male
				F = Female

		Text	categorical	Text31 Text42
		Task Order	categorical	V-B: Visual first, Blind second B-V: Blind first, Visual second
		Text Order	categorical	31-42: Text31 first, Text42 second 42-31: Text42 first, Text31 second
		Age	numeric	years
		Experience	numeric	years (working as a translator)
		Copy Time	numeric	seconds / 100 words
		Copy Effort	numeric	typed chars / target chars (%)
		Copy Errors	numeric	errors / 100 words

The statistical analysis in this section follows a mixed-effects linear regression model with repeated measures whenever a normal sample distribution can be assumed, or a generalised linear mixed model in all other cases. Mixed models have been advocated for as a better alternative for interpreting the data from naturalistic experiments as compared to other factorial designs such as ANOVA (see Balling 2008; Hvelplund 2011; Green et al. 2013). Its main advantages are the possibility to include as many independent variables as necessary, in order to check whether they have a significant effect on a particular dependent variable, the possibility to check the effect of the interaction between independent variables and the possibility to take into account random effects such as inter-subject variations. In the current study, the participants are included as random effects to account for repeated measures, which correspond to the several measurement instances of the same dependent variable (time, typing, errors), distributed over 56 data points (2 tasks × 4 suggestion types × 7 segments) for each participant.

The first step in the statistical analysis consists of checking for the potential main effects of individual independent variables on a dependent variable. In subsequent steps, the independent variables for which no significant effects are found are progressively removed and the model is run again with the remaining variables, and interaction effects are included in addition to the main effects. Reducing the number of variables helps prevent non-significant variables from hiding the effects of actually significant variables. Balling (2008: 186 ff.) illustrates this method with a model that first includes "all available predictors […] and then [is] reduced in step-wise fashion, reaching a model which only included significant predictors".

The statistical tests used in this thesis were run with IBM SPSS Statistics version 22 using the MIXED and GENLINMIXED functions. The choice of this software over other possibilities such as R or SAS was based on its user-friendliness.

As mentioned above, my experiment design includes independent variables that can be numeric or categorical. A linear regression model can become unstable and produce misleading results if there is high collinearity between covariates (numeric variables). A high collinearity effect was expected between Age and Experience, as in my sample these two covariates have a strong correlation ($r_p = 0.929$, $p < 0.001$). For this reason, the estimations provided by the model with both variables defined as numeric would not be valid. The solution adopted when running the tests was to include one covariate at a time. None of them was determined to produce significant main effects on any of the dependent variables. When looking for interaction effects, the numeric variable Experience was converted into a categorical variable with two levels: high experience (more than five years working as a translator) and low experience (less than five years). The division point for the two levels was defined as the integer that was closest to the median of the variable distribution.

In the following sections, I present the results obtained according to the statistical method just explained, looking at each of the dependent variables in turn.

5.4.1. Translation time

The first dependent variable to be analysed is Translation Time, which is measured in seconds per 100 words. Since the sample data for this variable is extremely right-skewed (see Figure 16), a logarithmic transformation was applied, resulting in the new histogram shown in Figure 17. This log-transformed data is used as the dependent variable in the linear regression model, assuming a normal distribution.

Figure 16. Sample distribution for Translation Time

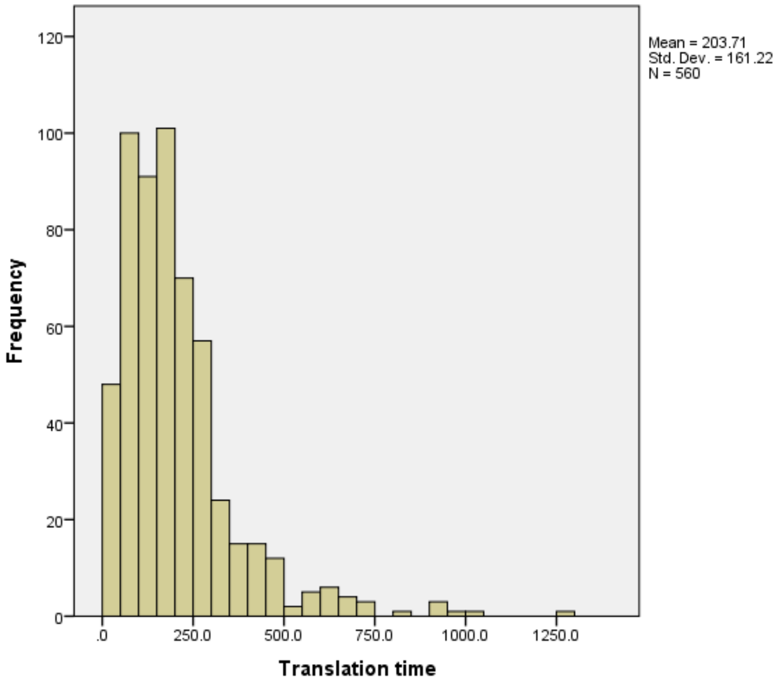
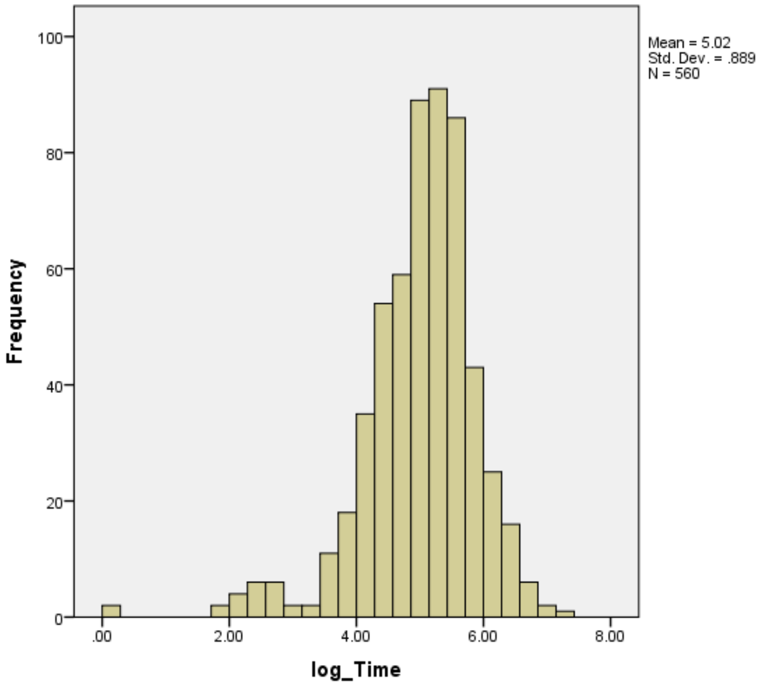


Figure 17. Sample distribution for Translation Time, after logarithmic transformation



The relevant independent variables are comprised of six factors – Task (metadata), Text, Suggestion Type, Gender, Task Order, Text Order –, and three covariates – Age, Experience and Copy Time (see Table 17). Copy Effort and Copy Errors are not included in the model at this stage, as they are not intuitively relevant as potential predictors for Translation Time. To avoid collinearity issues between the three covariates, the model was initially fit with each of them at a time, and configured to check only for main effects. The result of running this initial step, subdivided into three sub-steps, is indicated in Table 18 to Table 20. The tables indicate similar results, with significant main effects for Task (F = 14.18; p < 0.001), Suggestion Type (F = 79.08; p < 0.001) and Text (F = 3.912; p < 0.05). No significant main effects were detected for any of the other predictors, namely Task Order, Text Order, Gender, Age, Experience and Copy Time.

Table 18. Type III tests of fixed effects on Translation Time (log), including all factors and only Age as a covariate

Predictor	Numerator df	Denominator df	F	p*
Task	*1*	*545*	*14.18*	*0.000*
Text	*1*	*545*	*3.912*	*0.048*
Suggestion Type	*3*	*545*	*79.08*	*0.000*
Task Order	1	5	0.473	0.522
Text Order	1	5	0.074	0.797
Gender	1	5	1.129	0.337
Age	1	5	0.003	0.960

Table 19. Type III tests of fixed effects on Translation Time (log), including all factors and only Experience as a covariate

Predictor	Numerator df	Denominator df	F	p*
Task	*1*	*545*	*14.18*	*0.000*
Text	*1*	*545*	*3.912*	*0.048*
Suggestion Type	*3*	*545*	*79.08*	*0.000*
Task Order	1	5	0.550	0.491
Text Order	1	5	0.077	0.792
Gender	1	5	1.049	0.353
Experience	1	5	0.003	0.961

Table 20. Type III tests of fixed effects on Translation Time (log), including all factors and only Copy Time as a covariate

Predictor	Numerator df	Denominator df	F	p*
Task	*1*	*545*	*14.18*	*0.000*
Text	*1*	*545*	*3.912*	*0.048*
Suggestion Type	*3*	*545*	*79.08*	*0.000*
Task Order	1	5	0.533	0.498
Text Order	1	5	0.099	0.765
Gender	1	5	1.513	0.273
Copy Time	1	5	0.002	0.962

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

After running this first step, the model is fit again with the significant main effects from the first run and potentially relevant interaction effects between the factors. (Even if an independent variable does not appear as a significant main effect, it can still produce a significant interaction effect.) After multiple runs, where the non-significant effects are progressively eliminated, the final model is reached. It contains Translation Time (log-transformed) as the dependent variable, Task, Suggestion Type and Text as main effects, and Task × Suggestion Type, Task × Gender, Suggestion Type × Gender and Text × Experience (range) as interaction effects. The last factor – Experience (range) – corresponds to the previous covariate Experience, now converted into a categorical variable with two levels. Table 21 indicates the significant main and interaction effects for the selected factors.

Table 21. Type III tests of fixed effects on Translation Time (log), including significant main and interaction effects (final model)

Predictor	Numerator df	Denominator df	F	p*
Task	*1*	*537*	*32.71*	*0.000*
Suggestion Type	*3*	*537*	*104.6*	*0.000*
Text	*1*	*537*	*9.567*	*0.002*
Task × Suggestion Type	*3*	*537*	*38.80*	*0.000*
Task × Gender	*1*	*537*	*18.69*	*0.000*
Suggestion Type × Gender	*3*	*537*	*11.53*	*0.000*
Text × Experience (range)	*3*	*12*	*7.101*	*0.009*

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

The results indicate significant main effects for Task (F = 32.71; p < 0.001), Suggestion Type (F = 104.6; p < 0.001) and Text (F = 9.567; p = 0.002), and significant interaction effects between Task and Suggestion Type (F = 38.80; p < 0.001), Task and Gender (F = 18.69; p < 0.001), Suggestion Type and Gender (F = 11.53; p < 0.001), and Text and Experience (F = 7.101; p = 0.009). While the main effect of Task and the interaction effect between Task and Suggestion Type are directly related to my hypotheses (see sections 6.1.1 and 6.1.2), some of the remaining significant effects might indicate a pronounced difference between male and female participants (interaction effects for Gender) and even flaws in the research design (the significant main effect for Text). Each of the significant effects presented in Table 21 will be analysed in more detail in the following subsections.

5.4.1.1. Task (main effect)

The main effect of Task (F = 32.71; p < 0.001) is indicated in Table 22 through the estimated marginal means. The table presents the values for the means both in the logarithmic scale and in the original measurement for the dependent variable (seconds / 100 words). A mean difference of 179 - 125 = 54 seconds / 100 words was found between the two tasks, with a p-value smaller than 0.001. This corresponds to a difference of 43 percent between the two tasks. In other words, the statistical model estimates that *the translators spent 43 percent more time on the Blind task than on the Visual task on average*. It is worth noting that the results provided by the linear regression model might differ from the results obtained with other statistical tests comparing the means (such as a T-test), because the model takes into account the random variations between the participants and the interactions between different variables.

Table 22. Estimated marginal means for Translation Time, with Task as a main effect

Task	Mean (log)	Std. Error	Mean (secs / 100 words)
Visual (V)	4.837	0.123	125
Blind (B)	5.194	0.123	179

5.4.1.2. Suggestion Type (main effect)

The main effect of Suggestion Type on Translation Time has F = 104.6; p < 0.001. Figure 18 illustrates the estimated means for Translation Time with Suggestion Type as a main effect. Table 23 presents the values for the means both in the logarithmic scale and in the untransformed scale. Table 24 presents the results of pairwise comparisons based on the estimated marginal means. The results indicate that the mean differences are statistically significant between all the suggestion types, with p ≤ 0.003, except between Low Fuzzy Matches and Machine Translation, whose estimated mean difference of 13 seconds / 100 words is non-significant at p = 0.373.

Figure 18. Estimated means for Translation Time, with Suggestion Type as a main effect

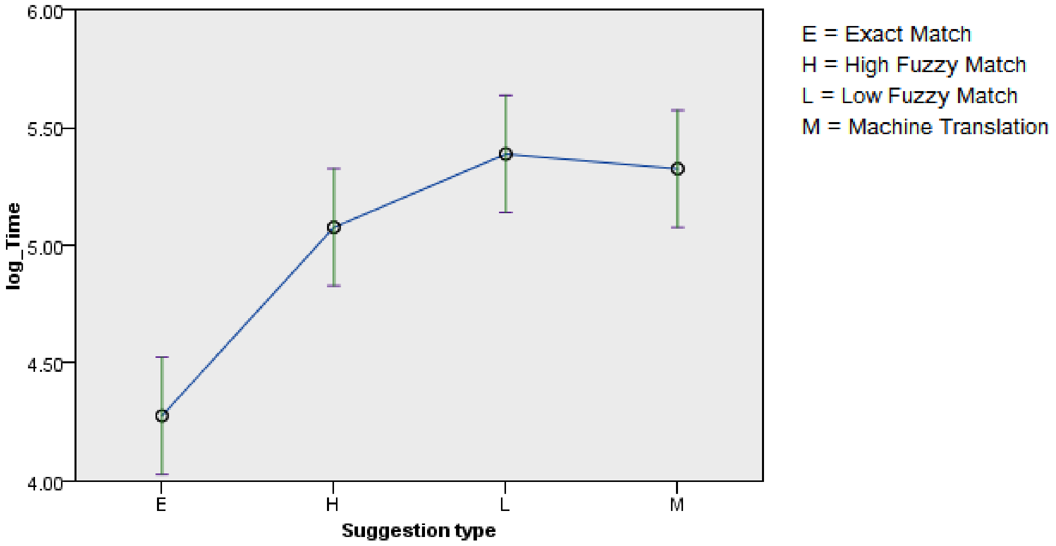


Table 23. Estimated marginal means for Translation Time, with Suggestion Type as a main effect

Suggestion Type	Mean (log)	Std. Error	Mean (secs / 100 words)
Exact Match (E)	4.276	0.127	71
High Fuzzy Match (H)	5.076	0.127	159
Low Fuzzy Match (L)	5.387	0.127	218
Machine Translation (M)	5.324	0.127	204

Table 24: Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with Suggestion Type as a main effect

Suggestion Types	Mean Difference	Mean Difference (%)	p*
E - H	*-88*	*-55%*	*0.000*
E - L	*-147*	*-67%*	*0.000*
E - M	*-133*	*-65%*	*0.000*
H - L	*-58*	*-27%*	*0.000*
H - M	*-45*	*-22%*	*0.003*
L - M	13	7%	0.373

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

Considering the differences between suggestion types, the results can be interpreted as follows:

- Translators spent 55% less time dealing with Exact Matches than with High Fuzzy Matches;
- Translators spent 65–67% less time dealing with Exact Matches than with Low Fuzzy Matches or Machine Translation;
- Translators spent 22–27% less time dealing with High Fuzzy Matches than with Low Fuzzy Matches or Machine Translation;

- Translators spent virtually the same time dealing with Low Fuzzy Matches and Machine Translation.

5.4.1.3. Text (main effect)

The main effect of Text (F = 9.567; p = 0.002) on Translation Time is indicated in Table 25 through the estimated marginal means. A mean difference of 24 seconds per 100 words was found between the two texts, with a p-value of 0.002. In other words, the model estimates that *the translators spent 17 percent more time on Text42 than on Text31 on average*. This unexpected difference between the two texts is discussed in section 6.2.4.

Table 25. Estimated marginal means for Translation Time, with Text as a main effect

Text	Mean (log)	Std. Error	Mean (secs / 100 words)
Text31	4.937	0.122	138
Text42	5.095	0.122	162

5.4.1.4. Task and Suggestion Type (interaction effect)

The interaction effect between Task and Suggestion Type on Translation Time has F = 38.80; p < 0.001. Figure 19 illustrates the estimated means for Translation Time with the interaction effect between the two factors. It shows that the effect of Suggestion Type is much higher in the Visual task than in the Blind task, as the means for the different suggestion types are much more spread apart in the Visual task. Table 26 presents the estimated values for the means per Task and Suggestion Type.

Figure 19. Estimated means for Translation Time, with the interaction effect between Task and Suggestion Type

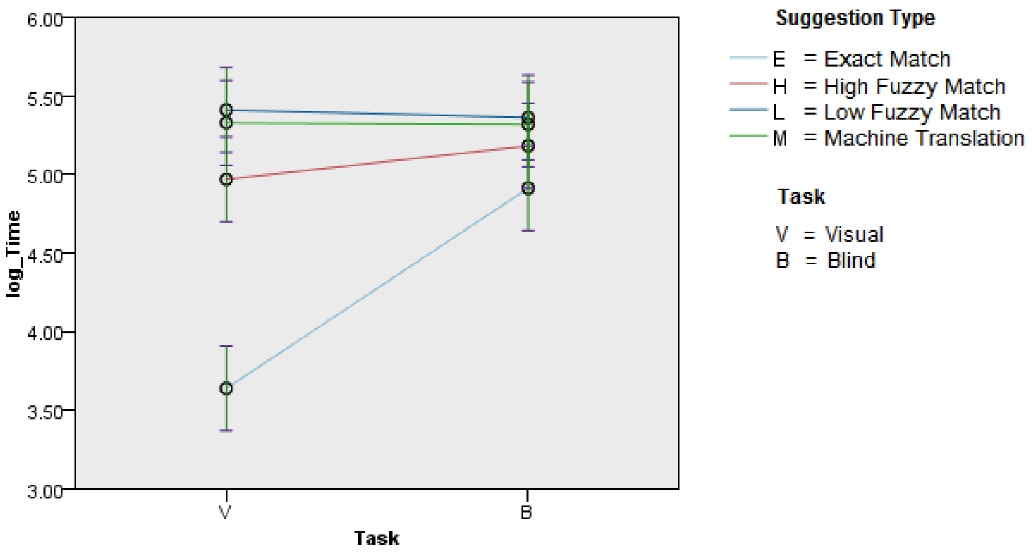


Table 26. Estimated marginal means for Translation Time, with the interaction effect between Task and Suggestion Type

Task	Suggestion Type	Mean (log)	Std. Error	Mean (secs/ 100 words)
Visual (V)	Exact Match (E)	3.638	0.138	37
	High Fuzzy Match (H)	4.969	0.138	143
	Low Fuzzy Match (L)	5.411	0.138	223
	Machine Translation (M)	5.329	0.138	205
Blind (B)	Exact Match (E)	4.913	0.138	135
	High Fuzzy Match (H)	5.183	0.138	177
	Low Fuzzy Match (L)	5.363	0.138	212
	Machine Translation (M)	5.319	0.138	203

Table 27 presents the results of pairwise comparisons, taking Task as the reference factor. The mean differences between the two tasks are significant for Exact Matches ($p < 0.001$) and High Fuzzy Matches ($p < 0.05$) and not significant for Low Fuzzy Matches and Machine Translation. These results can be read as:

- The translators spent 98 seconds per 100 words (265 percent) more time on average dealing with Exact Matches in the Blind task than in the Visual task;
- The translators spent 34 seconds per 100 words (24 percent) more time on average dealing with High Fuzzy Matches in the Blind task than in the Visual task;
- The time translators spent dealing with Low Fuzzy Matches or Machine Translation was not significantly different between the Blind task and the Visual task.

Table 27. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Task and Suggestion Type (Task as the reference factor)

Suggestion Type	Mean Difference (Blind - Visual)	Mean Difference (%)	p*
E	*98*	*265%*	*0.000*
H	*34*	*24%*	*0.046*
L	-10	-5%	0.650
M	-2	-1%	0.922

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

Table 28 presents the results of pairwise comparisons, when Suggestion Type is taken as the reference factor. In the Visual task, the difference in Translation Time is significant between all suggestion types, except between Low Fuzzy Matches and

Machine Translation. In the Blind task, there is still a significant difference in Translation Time between Exact Matches and the other suggestion types, but this difference is much lower, and the differences between the three other suggestion types are not statistically significant.

Table 28. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Task and Suggestion Type (Suggestion Type as the reference factor)

Task	Suggestion Types	Mean Difference	Mean Difference (%)	p*
Visual (V)	*E - H*	*-105.87*	*-74%*	*0.000*
	E - L	*-185.84*	*-83%*	*0.000*
	E - M	*-168.22*	*-82%*	*0.000*
	H - L	*-79.97*	*-36%*	*0.000*
	H - M	*-62.35*	*-30%*	*0.001*
	L - M	17.62	9%	0.412
Blind (B)	*E - H*	*-42.17*	*-24%*	*0.029*
	E - L	*-77.32*	*-36%*	*0.000*
	E - M	*-68.13*	*-34%*	*0.000*
	H - L	-35.15	-17%	0.216
	H - M	-25.96	-13%	0.346
	L - M	9.18	5%	0.660

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

The results in the current section refine the findings presented in section 5.4.1.1, when the tasks were considered regardless of suggestion types, as well as those presented in section 5.4.1.2, when the suggestion types were considered regardless of task.

5.4.1.5. Task and Gender (interaction effect)

The interaction effect between Task and Gender on Translation Time was also determined to be significant, with F = 18.69; $p < 0.001$. Figure 20 illustrates the estimated means for Translation Time with the interaction effect between the two factors. Although the figure suggests that translation times are consistently lower for women than for men in both tasks, Gender was not determined to be a significant main effect (see Table 18 to Table 20). The effect of Gender is much higher in the Visual task than in the Blind task, as the means for men and women are much more spread apart in the Visual task.

Table 29 presents the estimated values for the means per Task and Gender. Table 30 presents the results of the post-hoc analysis, taking Task as the reference factor for pairwise comparisons. The results indicate that the mean differences between the two tasks are statistically significant for women ($p < 0.001$), but not for men.

Figure 20. Estimated means for Translation Time, with the interaction effect between Task and Gender

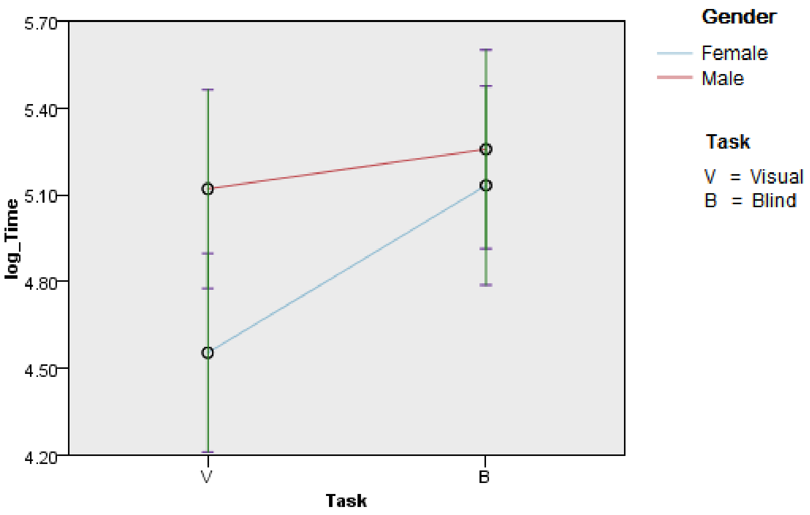


Table 29. Estimated marginal means for Translation Time, with the interaction effect between Task and Gender

Task	Gender	Mean (log)	Std. Error	Mean (secs / 100 words)
Visual (V)	Female (F)	4.554	0.175	94
	Male (M)	5.120	0.175	166
Blind (B)	Female (F)	5.132	0.175	168
	Male (M)	5.257	0.175	191

Table 30. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Task and Gender (Task as the reference factor)

Gender	Mean Difference (Blind - Visual)	Mean Difference (%)	p*
Female (F)	*74*	*79%*	*0.000*
Male (M)	25	15%	0.090

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

Table 31 shows that when Gender is taken as the reference factor for the pairwise comparisons, the mean differences between men and women in any given task are never significant, which coincides with the fact that Gender is not a significant main effect.

Table 31. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Task and Gender (Gender as the reference factor)

Task	Mean Difference (Male - Female)	Mean Difference (%)	p*
Visual (V)	72	77%	0.054
Blind (B)	23	13%	0.628

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons.

The overall results of the interaction effect can be summarised as follows: On average, women participants spent significantly less time on the Visual task than on the Blind task.

5.4.1.6. Suggestion Type and Gender (interaction effect)

The interaction effect between Suggestion Type and Gender on Translation Time was also determined to be significant, with F = 11.53; p < 0.001. Figure 21 illustrates the estimated means for Translation Time with the interaction effect between the two factors. Although the figure suggests that translation times are consistently lower for women than for men across all suggestion types, it is worth recalling that Gender was not determined to have a significant main effect (see Table 18 to Table 20).

Figure 21. Estimated means for Translation Time, with the interaction effect between Suggestion Type and Gender

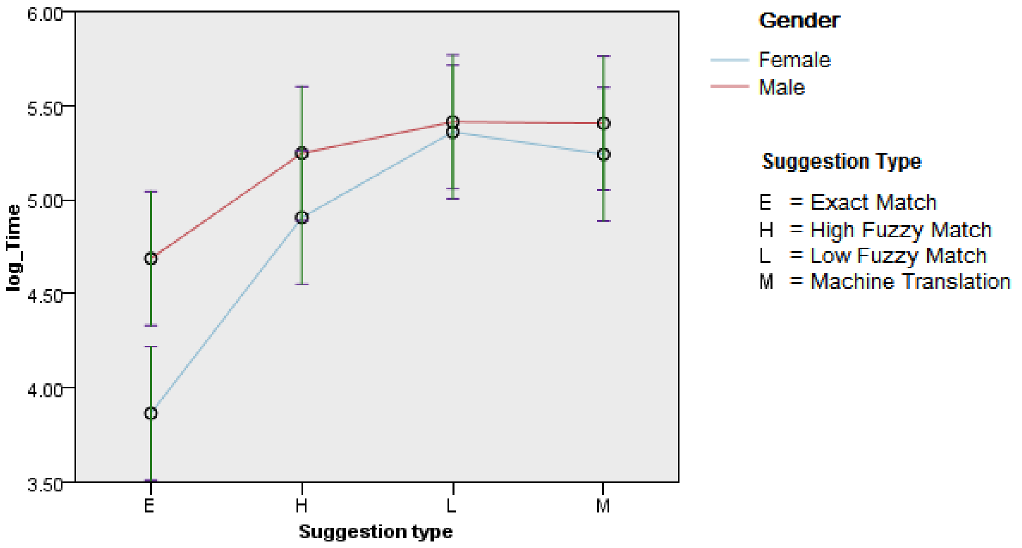


Table 32 presents the estimated values for the means per Suggestion Type and Gender. Table 33 presents the results of the post-hoc analysis, taking Suggestion Type as the reference factor for pairwise comparisons. The results are qualitatively similar to those presented in section 5.4.1.2, where Suggestion Type was analysed as a main effect, except that for men the mean differences are not statistically significant between High Fuzzy Matches and Low Fuzzy Matches and between High Fuzzy Matches and Machine Translation, in addition to the non-statistically significant difference between Low Fuzzy Matches and Machine Translation, which is common to both genders.

Table 32. Estimated marginal means for Translation Time, with the interaction effect between Suggestion Type and Gender

Suggestion Type	Gender	Mean (log)	Std. Error	Mean (secs / 100 words)
Exact Match (E)	Female (F)	3.864	0.181	47
	Male (M)	4.687	0.181	108
High Fuzzy Match (H)	Female (F)	4.905	0.181	134
	Male (M)	5.247	0.181	189
Low Fuzzy Match (L)	Female (F)	5.360	0.181	212
	Male (M)	5.414	0.181	224
Machine Translation (M)	Female (F)	5.242	0.181	188
	Male (M)	5.407	0.181	222

Table 33. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Suggestion Type and Gender (Suggestion Type as the reference factor)

Gender	Suggestion Types	Mean Difference	Mean Difference (%)	p*
Female (F)	*E - H*	*-87*	*-65%*	*0.000*
	E - L	*-165*	*-78%*	*0.000*
	E - M	*-141*	*-75%*	*0.000*
	H - L	*-78*	*-37%*	*0.000*
	H - M	*-54*	*-29%*	*0.002*
	L - M	24	13%	0.236
Male (M)	*E - H*	*-81*	*-43%*	*0.000*
	E - L	*-116*	*-52%*	*0.000*
	E - M	*-114*	*-52%*	*0.000*
	H - L	-35	-15%	0.284
	H - M	-33	-15%	0.284
	L - M	2	1%	0.942

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

Table 34 shows the results of pairwise comparisons when Gender is taken as the reference factor. This indicates that the mean differences between men and women are only significant for Exact Matches.

Table 34. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Suggestion Type and Gender (Gender as the reference factor)

Suggestion Type	Mean Difference (Male - Female)	Mean Difference (%)	p*
E	*61*	131%	*.011*
H	55	41%	.219
L	12	6%	.839
M	34	18%	.538

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

The results of the interaction effect in this section can be summarised as follows:

- Both men and women spent significant less time translating Exact Matches than any other type of translation suggestions.
- Women spent significant less time translating High Fuzzy Matches than Low Fuzzy Matches or Machine Translation.
- For both men and women, there was no significant difference between the time spent translating Low Fuzzy Matches and Machine Translation.
- Women spent significantly less time than men when translating Exact Matches.

5.4.1.7. Text and Experience (interaction effect)

The interaction effect between Text and Experience on Translation Time is the last one to have been determined as significant, with F = 7.10; p = 0.009. Figure 22 illustrates the estimated means for Translation Time with the interaction effect between the two factors. The values for the means are then presented in Table 35.

Figure 22. Estimated means for Translation Time, with the interaction effect between Text and Experience

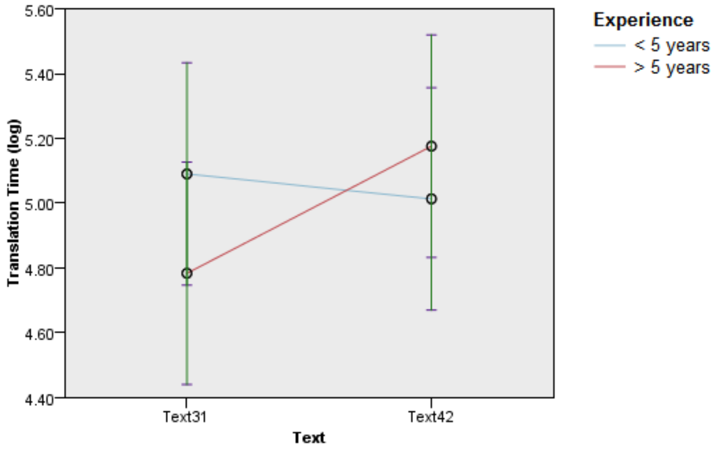


Table 35. Estimated marginal means for Translation Time, with the interaction effect between Text and Experience

Text	Experience	Mean (log)	Std. Error	Mean (secs / 100 words)
Text31	<5	5.090	0.175	161
	>5	4.783	0.175	118
Text42	<5	5.013	0.175	149
	>5	5.176	0.175	176

The results of the pairwise comparisons are presented in Table 36 (Text as the reference factor) and Table 37 (Experience as the reference factor). Table 36 indicates that the more experienced translators (more than five years of work experience) spent significantly more time (58 seconds per 100 words, or 49 percent) translating Text42 than translating Text31. No other mean differences were determined to be statistically significant. These findings refine those presented in section 5.4.1.3, where Text had been identified as a significant main effect on Translation Time.

Table 36. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Text and Experience (Text as the reference factor)

Experience	Mean Difference (Text42 - Text31)	Mean Difference (%)	p*
< 5 years	-12	-7%	0.341
> 5 years	*58*	49%	*0.000*

Table 37. Pairwise comparisons of estimated marginal means for Translation Time (seconds / 100 words), with the interaction effect between Text and Experience (Experience as the reference factor)

Text	Mean Difference (>5 - <5)	Mean Difference (%)	p*
Text31	-43	-27%	0.222
Text42	27	18%	0.517

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

5.4.2. Typing effort (preliminary considerations)

The second dependent variable to be analysed is Typing Effort, measured as a ratio between the number of characters typed by the translator and the total number of characters in the final target segment. Similarly to Translation Time, the data for Typing Effort does not follow a normal distribution. As can be seen in Figure 23, it is also extremely right-skewed, with a high peak at zero (many segments required no edits). Unlike what happened with Translation Time, however, applying a logarithmic transformation does not normalise the distribution, since the concentration of data points around zero remains, as indicated in Figure 24. In this case, I chose to analyse the data using a two-step method.

Figure 23. Sample distribution for Typing Effort

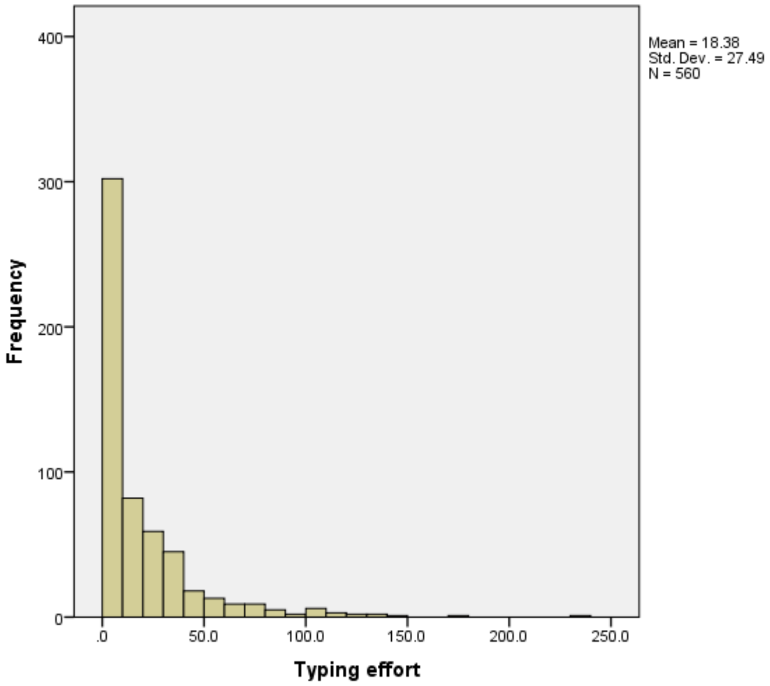
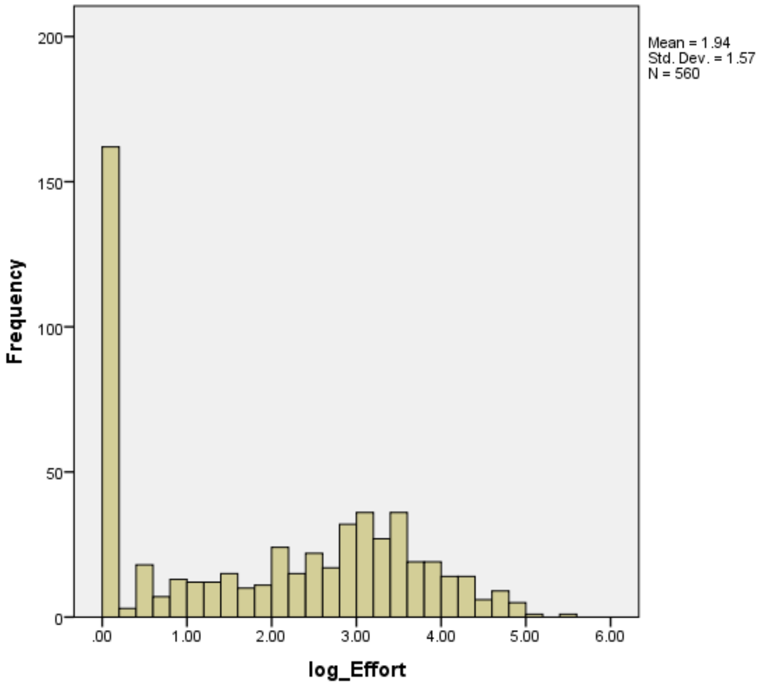


Figure 24. Sample distribution for Typing Effort, after logarithmic transformation



The first step consists in transforming Typing Effort into a categorical variable with two levels, where 0 corresponds to zero typing effort and 1 corresponds to any value greater than zero. The dependent variable thus transformed is selected as the target for a generalised linear mixed model with a binomial distribution. The second step consists in eliminating the data points that were equal to zero and using the remaining data points as the dependent variable in the linear mixed-effects model. In the following sections, the results of the two-step analysis will be presented.

5.4.3. Typing effort (binary)

In the first step, a generalised linear mixed model with a binomial distribution tests for main effects with all the factors (Task, Text, Suggestion Type, Gender, Task Order, and Text Order) and covariates (Age, Experience and Copy Effort) included as independent variables (see Table 17).

Following the same procedure described for Translation Time (see section 5.4.1), the covariates are included one by one, to avoid collinearity issues, and then all the non-significant main effects are eliminated progressively. In all these iterations, only Suggestion Type is determined to have a significant main effect (F ≈ 43.2; p < 0.001). Finally, the model is configured to test for interaction effects. Table 38 presents the significant main effects and interaction effects obtained with the final model that was determined using this process.

Table 38. Type III tests of fixed effects on Typing Effort (binary), including significant main and interaction effects (final model)

Predictor	Numerator df	Denominator df	F	p*
Task	*1*	*551*	*6.266*	*0.013*
Suggestion Type	*3*	*551*	*41.16*	*0.000*
Task × Suggestion Type	*3*	*551*	*8.665*	*0.000*

* Rows in italics indicate significant results (α = 0.05).

The results indicate significant main effects for Task (F = 6.266; p = 0.013) and Suggestion Type (F = 41.16; p < 0.001), and significant interaction effects between Task and Suggestion Type (F = 8.665; p < 0.001). The significant effects presented in Table 38 will be analysed in the following subsections.

5.4.3.1. Task (main effect)

The main effect of Task (F = 6.266; p = 0.013) is indicated graphically in Figure 25 and through the estimated marginal means in Table 39. It is worth recalling that the dependent variable Typing Effort was converted into a binary variable, where 0 corresponds to no edits and 1 corresponds to any value greater than zero edits. The results can thus be interpreted as follows: 83 percent of segments in the Visual task required some degree of editing (and 17 percent required no editing at all), while only 70.3 percent of segments in the Blind task required some degree of editing (and 29.7 percent required no editing at all). The model indicates that the mean difference of 0.126 between the two tasks is statistically significant, with a p-value of 0.008. This corresponds to a difference of 18 percent between the two tasks. In other words, the statistical model estimates that *the translators edited 18 percent more segments in the Visual task than in the Blind task on average*.

Figure 25. Estimated means for Typing Effort (binary), with Task as a main effect

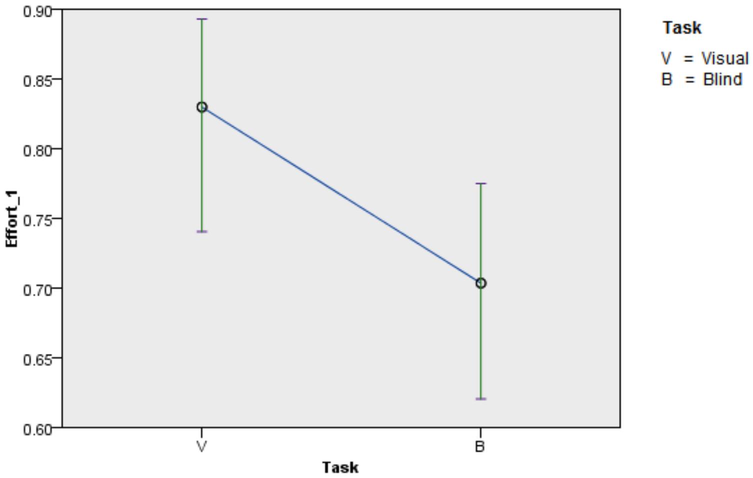


Table 39. Estimated marginal means for Typing Effort (binary), with Task as a main effect

Task	Mean	Std. Error
Visual (V)	0.830	0.039
Blind (B)	0.703	0.040

5.4.3.2. Suggestion Type (main effect)

The main effect of Suggestion Type on Typing Effort (F = 41.16; p < 0.001) is indicated graphically in Figure 26 and through the estimated marginal means in Table 40. Exact Matches stand out as the suggestion type that required editing the least frequently (only 25.7 percent of segments with an Exact Match as the translation suggestion required some degree of editing, compared to approximately 81–94 percent for the other suggestion types). As indicated in Table 41, the mean differences are statistically significant between all suggestion types, except between High Fuzzy Matches and Machine Translation.

Figure 26. Estimated means for Typing Effort (binary), with Suggestion Type as a main effect

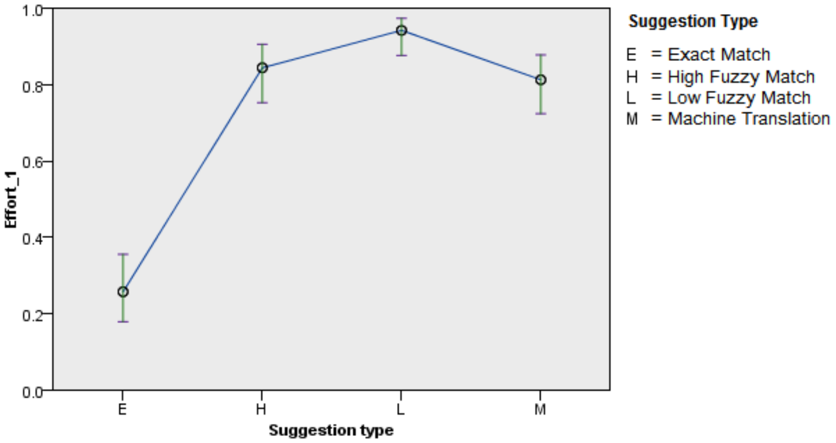


Table 40. Estimated marginal means for Typing Effort (binary), with Suggestion Type as a main effect

Suggestion Type	Mean	Std. Error
Exact Match (E)	0.257	0.045
High Fuzzy Match (H)	0.844	0.039
Low Fuzzy Match (L)	0.942	0.023
Machine Translation (M)	0.813	0.039

Table 41. Pairwise comparisons of estimated marginal means for Typing Effort (binary), with Suggestion Type as a main effect

Suggestion Types	Mean Difference	Mean Difference (%)	p*
E - H	*-0.587*	*-70%*	*.000*
E - L	*-0.685*	*-73%*	*.000*
E - M	*-0.556*	*-68%*	*.000*
H - L	*-0.098*	*-10%*	*.044*
H - M	0.031	4%	.527
L - M	*0.129*	*16%*	*.008*

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

Considering the differences between suggestion types, the results can be interpreted as:

- Segments that have an Exact Match as the translation suggestion are edited 70% less frequently than segments with a High Fuzzy Match, 73% less frequently than segments with a Low Fuzzy Match and 68% less frequently than segments with Machine Translation;
- Segments that have a High Fuzzy Match as the translation suggestion are edited 10% less frequently than segments with a Low Fuzzy Match but as frequently as segments with Machine Translation (non-significant difference);

- Segments that have a Low Fuzzy Match as the translation suggestion are edited 16% more frequently than segments with Machine Translation.

5.4.3.3. Task and Suggestion Type (interaction effect)

Task and Suggestion Type were determined to have a significant interaction effect (F = 8.66; p < 0.001) on Typing Effort. Figure 27 shows a chart for the estimated means and confidence intervals for this effect. The chart suggests that Exact Matches (E), represented by the bottom-most line, require more edits in the Blind task (B) than in the Visual task (V), while all the other suggestion types require more edits in the Visual task. The exact mean values and standard errors are presented in Table 42, while the mean differences and respective significances are presented in Table 43 and Table 44.

Figure 27. Estimated means for Typing Effort (binary), with the interaction effect between Task and Suggestion Type

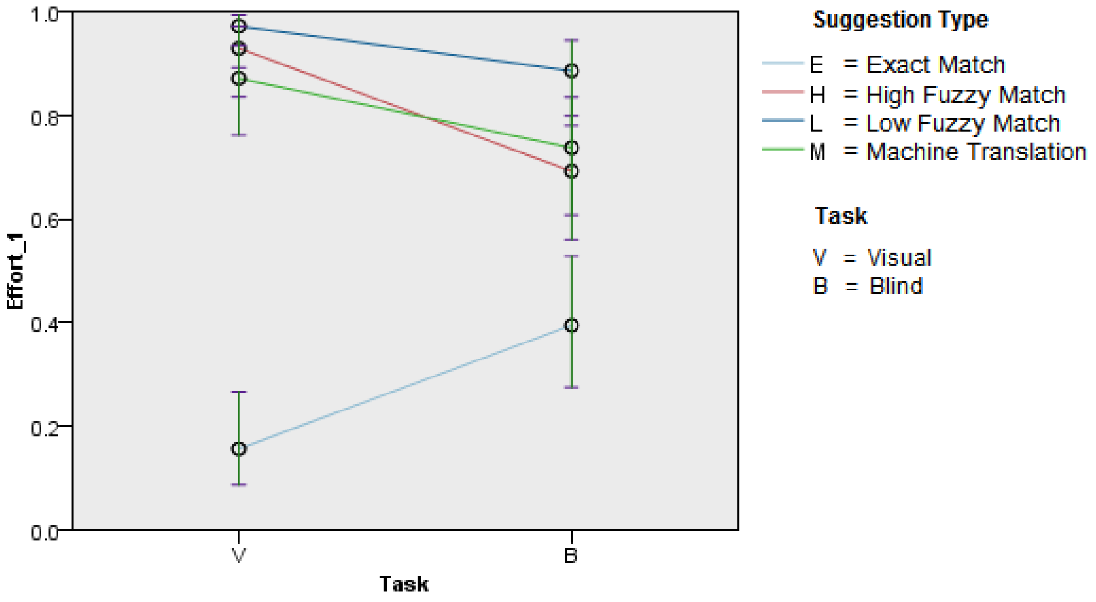


Table 42. Estimated marginal means for Typing Effort (binary), with the interaction effect between Task and Suggestion Type

Task	Suggestion Type	Mean	Std. Error
Visual (V)	Exact Match (E)	0.156	0.045
	High Fuzzy Match (H)	0.929	0.032
	Low Fuzzy Match (L)	0.972	0.020
	Machine Translation (M)	0.871	0.043
Blind (B)	Exact Match (E)	0.394	0.066
	High Fuzzy Match (H)	0.692	0.062
	Low Fuzzy Match (L)	0.885	0.040
	Machine Translation (M)	0.737	0.059

The results in Table 43 indicate that the mean differences between the two tasks are significant for Exact Matches (p = 0.001) and High Fuzzy Matches (p < 0.001) and on the verge of significance for Low Fuzzy Matches (p = 0.051) and Machine Translation (p = 0.050). Looked at from another perspective, *translation metadata (present in the Visual task) reduce typing effort only for Exact Matches and have no effect or are even detrimental for the other types of translation suggestions*. (Translators made fewer changes or as many changes to the translation suggestions when they did not know the type of suggestion they were editing, except for Exact Matches.)

The results in Table 44 indicate that the mean differences between Exact Matches and the three other suggestion types are significant in both tasks, while the difference between High Fuzzy Matches and Low Fuzzy Matches is significant in the Blind task. The results can be summarised as follows:

- In the Visual task, Exact Matches require the fewest edits, followed by High Fuzzy Matches, Low Fuzzy Matches and Machine Translations (no significant difference between the three).
- In the Blind task, Exact Matches require the fewest edits, followed by High Fuzzy Matches, followed by Machine Translations and Low Fuzzy Matches (no significant difference between the two).

Table 43. Pairwise comparisons of estimated marginal means for Typing Effort (binary), with the interaction effect between Task and Suggestion Type (Task as the reference factor)

Suggestion Type	Mean Difference (Blind - Visual)	Mean Difference (%)	p*
E	*0.238*	*153%*	*0.001*
H	*-0.237*	*-26%*	*0.000*
L	-0.086	-9%	0.051
M	-0.134	-15%	0.050

Table 44. Pairwise comparisons of estimated marginal means for Typing Effort (binary), with the interaction effect between Task and Suggestion Type (Suggestion Type as the reference factor)

Task	Suggestion Types	Mean Difference	Mean Difference (%)	p*
Visual (V)	*E - H*	*-0.773*	*-83%*	*0.000*
	E - L	*-0.816*	*-84%*	*0.000*
	E - M	*-0.715*	*-82%*	*0.000*
	H - L	-0.043	-4%	0.491
	H - M	0.058	7%	0.491
	L - M	0.101	12%	0.088
Blind (B)	*E - H*	*-0.298*	*-43%*	*0.001*
	E - L	*-0.491*	*-55%*	*0.000*
	E - M	*-0.343*	*-47%*	*0.000*
	H - L	*-0.194*	*-22%*	*0.017*
	H - M	-0.045	-6%	0.565
	L - M	0.149	20%	0.055

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

5.4.4. Typing effort (non-zero cases)

The second step for analysing the non-normally distributed data for Typing Effort consisted in eliminating the data points that were equal to zero and using the log-transformed data for the remaining data points (see Figure 24). From a total of 560 data points, 162 were equal to zero and were eliminated and the remaining 398 constituted the new dataset. The resulting numeric variable was used as the dependent variable in the linear mixed-effects model, assuming a normal distribution. Following the same procedure described in the previous sections, the model tests for main effects and then for interaction effects, and the non-significant effects are eliminated progressively. As opposed to the first step of the analysis, the statistical model found no significant main effect for Task in this second step, with the log-transformed non-zero data points.

The results of the final model are presented in Table 45. They indicate significant main effects for Suggestion Type (F = 30.30; p < 0.001), and significant interaction effects between Task and Suggestion Type (F = 3.224; p = 0.023). The non-significant main effect of Task is also included in the table, because this result is necessary for testing my second hypothesis (see section 6.1.3). The significant results presented in Table 45 will be analysed in the following subsections.

Table 45. Type III tests of fixed effects on Typing Effort (non-zero, log), including significant main and interaction effects (final model)

Predictor	Numerator df	Denominator df	F	p*
Task	1	389	2.430	0.120
Suggestion Type	*3*	*384*	*30.30*	*0.000*
Task × Suggestion Type	*3*	*386*	*3.224*	*0.023*

* Rows in italics indicate significant results (α = 0.05).

5.4.4.1. Suggestion Type (main effect)

The main effect of Suggestion Type on Typing Effort (F = 30.30; p < 0.001) is indicated graphically in Figure 28 and through the estimated marginal means in Table 46. When comparing these results with the ones presented in section 5.4.3.2 (Figure 26 and Table 40), we should bear in mind that:

- we have now recovered the original meaning of the variable (after undoing the logarithmic transformation), which is a ratio between the number of characters typed by the translator and the total number of characters in the final target segment, indicated as a percentage;
- we have removed from the analysis all those segments for which the typing effort was zero, i.e. for which absolutely no character was changed in the translation suggestion.

For example, the results for Exact Matches in Table 46 indicate that, for those segments where some change was made to the translation suggestion, the translators typed only 4.3 percent of the characters required to produce the final translation. (The remaining 95.7 percent of the characters were already in the translation suggestion.)

Figure 28. Estimated means for Typing Effort (non-zero, log), with Suggestion Type as a main effect

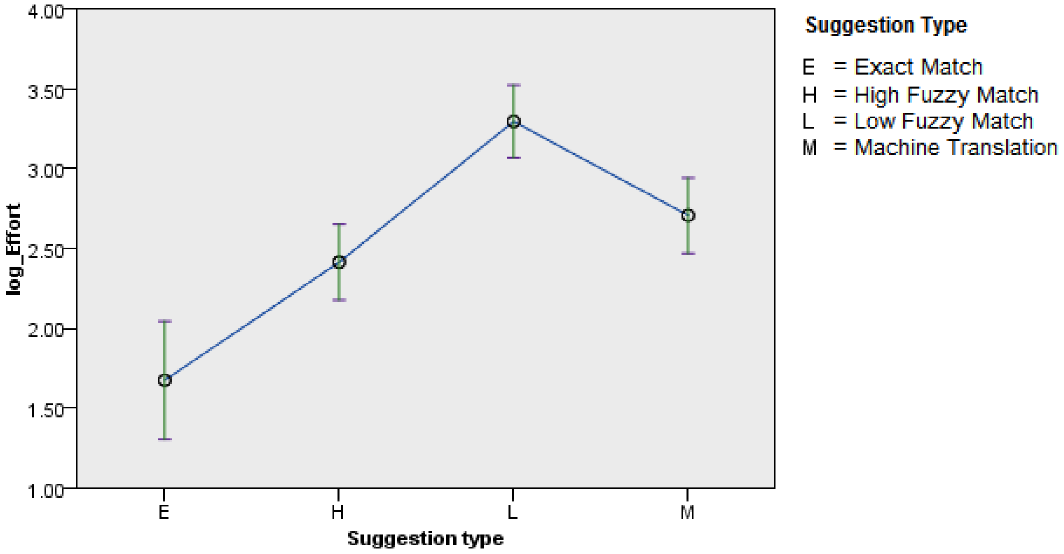


Table 46. Estimated marginal means for Typing Effort (non-zero), with Suggestion Type as a main effect

Suggestion Type	Mean (log)	Std. Error	Mean (%)
Exact Match (E)	1.673	0.188	4.3
High Fuzzy Match (H)	2.414	0.121	10.2
Low Fuzzy Match (L)	3.296	0.115	26.0
Machine Translation (M)	2.707	0.120	14.0

Table 47 indicates that the mean differences are statistically significant between all suggestion types, except between High Fuzzy Matches and Machine Translation. Qualitatively, these results match those presented in Table 41 for the binary variable, when all segments are considered.

Table 47. Pairwise comparisons of estimated marginal means for Typing Effort (non-zero), with Suggestion Type as a main effect

Suggestion Types	Mean Difference	Mean Difference (%)	p*
E - H	*-5.9*	*-57%*	*0.001*
E - L	*-21.7*	*-83%*	*0.000*
E - M	*-9.7*	*-69%*	*0.000*
H - L	*-15.8*	*-61%*	*0.000*
H - M	-3.8	-27%	0.170
L - M	*12.0*	*86%*	*0.000*

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

The results in Table 47 indicate that, regardless of the task (i.e. regardless of the presence or absence of metadata):

- Exact Matches required less editing than High Fuzzy Matches, Low Fuzzy Matches or Machine Translation;
- High Fuzzy Matches required less editing than Low Fuzzy Matches, but as much editing as Machine Translation (non-significant difference);
- Low Fuzzy Matches required more editing than Machine Translation.

5.4.4.2. Task and Suggestion Type (interaction effect)

Task and Suggestion Type were determined to have a significant interaction effect (F = 3.224; p = 0.023) on Typing Effort. Figure 29 shows a chart for the estimated means and confidence intervals for this effect. Similarly to Figure 27 in section 5.4.3.3, the chart suggests that Exact Matches (E) require more edits in the Blind task (B) than in the Visual task (V), while all the other suggestion types require more edits in the Visual task. Due to the elimination of the segments with zero typing effort, however, the slopes of the lines for each suggestion type are noticeably different from those in Figure 27. The exact mean values and standard errors are presented in Table 48. The mean differences and respective significances are presented in Table 49 and Table 50.

Figure 29. Estimated means for Typing Effort (non-zero, log), with the interaction effect between Task and Suggestion Type

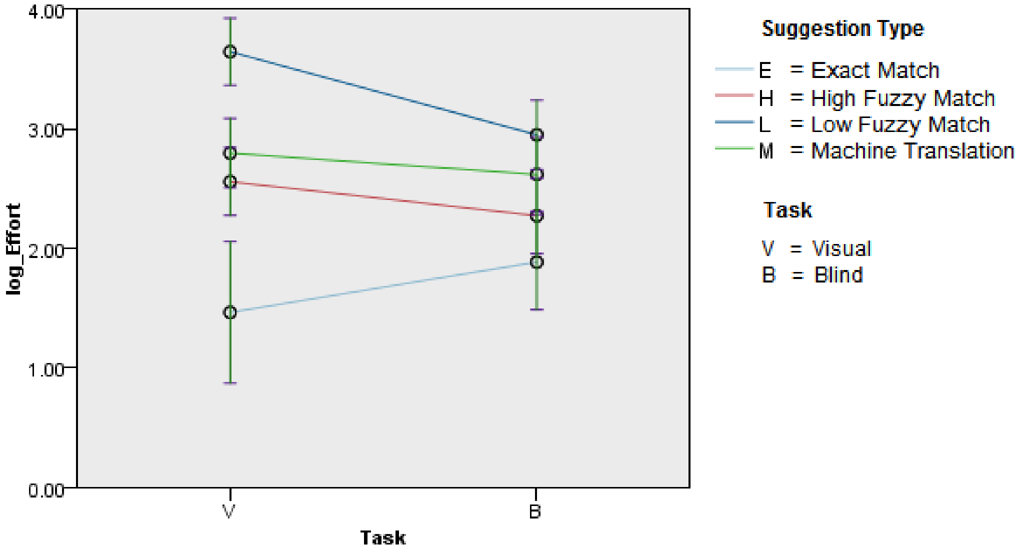


Table 48. Estimated marginal means for Typing Effort (non-zero), with the interaction effect between Task and Suggestion Type

Task	Suggestion Type	Mean (log)	Std. Error	Mean (%)
Visual (V)	Exact Match (E)	1.463	0.302	3.3
	High Fuzzy Match (H)	2.557	0.145	11.9
	Low Fuzzy Match (L)	3.643	0.142	37.2
	Machine Translation (M)	2.796	0.148	15.4
Blind (B)	Exact Match (E)	1.884	0.203	5.6
	High Fuzzy Match (H)	2.271	0.161	8.7
	Low Fuzzy Match (L)	2.949	0.147	18.1
	Machine Translation (M)	2.618	0.157	12.7

Table 49 presents the results of comparing the mean Typing Effort between the Visual and the Blind tasks for a given Suggestion Type. Although the signs of the mean differences in Table 49 are similar to those in Table 43, the statistical significances do not coincide. In the first step of the analysis the mean differences between the two tasks were found to be significant for Exact Matches (p = 0.001) and High Fuzzy Matches (p < 0.001); now they are significant only for Low Fuzzy Matches (p < 0.001). In this case, removing the segments for which no typing effort was required – as we have done in the current section – produced different results.

Table 49. Pairwise comparisons of estimated marginal means for Typing Effort (non-zero), with the interaction effect between Task and Suggestion Type (Task as the reference factor)

Suggestion Type	Mean Difference (Blind - Visual)	Mean Difference (%)	p*
E	2.3	68%	0.230
H	-3.2	-27%	0.129
L	*-19.1*	*-51%*	*0.000*
M	-2.7	-17%	0.344

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

Table 50 presents the results of comparing the mean Typing Effort between different suggestion types within a given task. In the Visual task, Exact Matches require less Typing Effort than any of the other suggestion types, High Fuzzy Matches require less Typing Effort than Low Fuzzy Matches, and Machine Translations require less Typing Effort than Low Fuzzy Matches. In the Blind task, the difference between Exact

Matches and High Fuzzy Matches ceases to be significant, as does the difference between Low Fuzzy Matches and Machine Translations.

Table 50. Pairwise comparisons of estimated marginal means for Typing Effort (non-zero), with the interaction effect between Task and Suggestion Type (Suggestion Type as the reference factor)

Task	Suggestion Types	Mean Difference	Mean Difference (%)	p*
Visual (V)	*E - H*	*-8.6*	*-72%*	*0.004*
	E - L	*-33.9*	*-91%*	*0.000*
	E - M	*-12.1*	*-78%*	*0.000*
	H - L	*-25.3*	*-68%*	*0.000*
	H - M	-3.5	-23%	0.180
	L - M	*21.8*	*142%*	*0.000*
Blind (B)	E - H	-3.1	-36%	0.232
	E - L	*-12.5*	*-69%*	*0.000*
	E - M	*-7.1*	*-56%*	*0.007*
	H - L	*-9.4*	*-52%*	*0.002*
	H - M	-4.0	-32%	0.232
	L - M	5.4	42%	0.232

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

As a general summary:

- In the Visual task, Exact Matches require the least Typing Effort, followed by High Fuzzy Matches and Machine Translations (no significant difference between the two), and followed by Low Fuzzy Matches.
- In the Blind task, Exact Matches and High Fuzzy Matches require the least Typing Effort (no significant difference between the two), followed by Low Fuzzy Matches and Machine Translations (no significant difference between the two).

These results are mostly similar to those presented in section 5.4.3.3, with some differences in the conclusions as far the statistical significances are concerned, due to the different treatment given to the dependent variable.

5.4.5. Error score (preliminary considerations)

Error score is the third and last of the dependent variables under study in this thesis. It is measured as the number of errors per 100 words of source text, based on a quality assessment done by human reviewers. Similarly to the two previous variables, the data for Error Score do not follow a normal distribution (Figure 30). Like what happens with Typing Effort, there is also a high peak at zero (in many segments, translators made no errors) and applying a logarithmic transformation does not normalise the distribution, since the peak at zero remains, as indicated in Figure 31. Therefore, as in the previous case, the variable will be analysed using a two-step method.

Figure 30. Sample distribution for Error Score

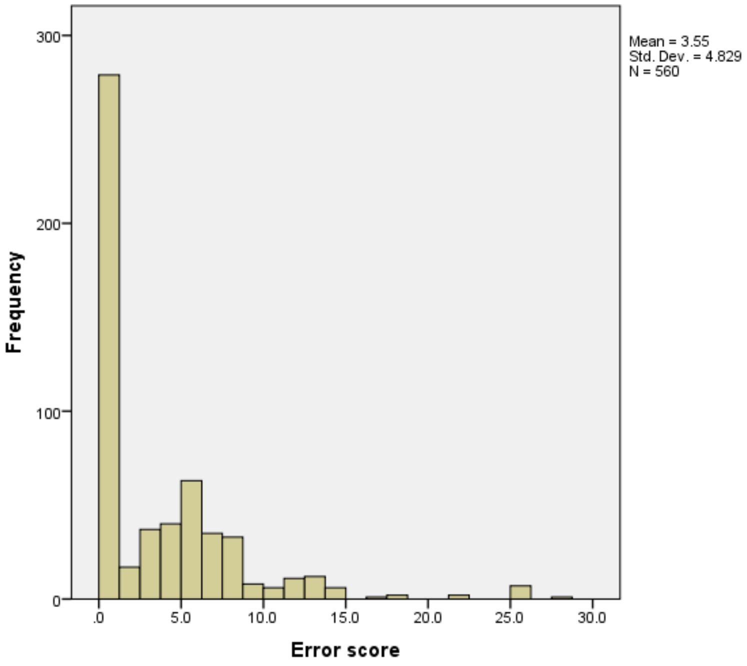
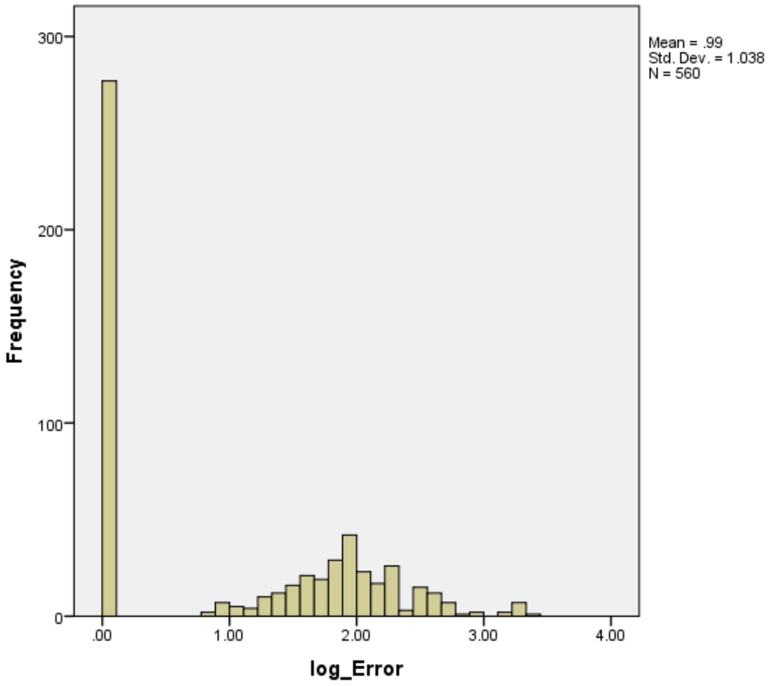


Figure 31. Sample distribution for Error Score, after logarithmic transformation



5.4.6. Error score (binary)

The first step consists in transforming Error Score into a categorical variable with two levels, where 0 corresponds to zero errors and 1 corresponds to any value greater than zero. The dependent variable thus transformed is selected as the target for a generalised linear mixed model with a binomial distribution. Initially, the model tests for main effects with all the factors (Task, Text, Suggestion Type, Gender, Task Order and Text Order) and covariates (Age, Experience and Copy Errors) included as independent variables (see Table 17). Following the same procedure described for Translation Time (section 5.4.1) and Typing Effort (sections 5.4.3 and 5.4.4), the covariates are included one by one, to avoid collinearity issues, and then all the non-significant effects are eliminated progressively.

In all these iterations, only Suggestion Type is determined to have a significant effect. The model is then configured to test for interaction effects, but this time no significant interactions are found. Since my hypotheses deal with the effects of Task (metadata) and its interaction with Suggestion Type, these effects are maintained in the final model, even though they are not found to be statistically significant. Table 51 presents the final statistical model with the significant main effect of Suggestion Type, the non-significant main effect of Task and the non-significant interaction effect between Task and Suggestion Type.

Table 51. Type III tests of fixed effects on Error Score (binary), including a significant main effect (final model)

Predictor	Numerator df	Denominator df	F	p*
Task	1	552	0.983	0.322
Suggestion Type	*3*	*552*	*9.208*	*0.000*
Task × Suggestion Type	3	552	0.138	0.937

* Rows in italics indicate significant results (α = 0.05).

5.4.6.1. Suggestion Type (main effect)

The significant effect of Suggestion Type (F = 9.208; p < 0.001) is indicated graphically in Figure 32 and through the estimated marginal means in Table 52. It is worth recalling that the dependent variable Error Score was converted into a binary variable, where 0 corresponds to no errors and 1 corresponds to any value greater than zero errors. The results can thus be interpreted as follows: The translators made at least one error in 41 percent of the segments that had an Exact Match as the translation suggestion, in 65 percent of the segments with a High Fuzzy Match, in 37 percent of the segments with a Low Fuzzy Match, and in 59 percent of the segments with a Machine Translation suggestion.

Figure 32. Estimated means for Error Score (binary), with Suggestion Type as a main effect

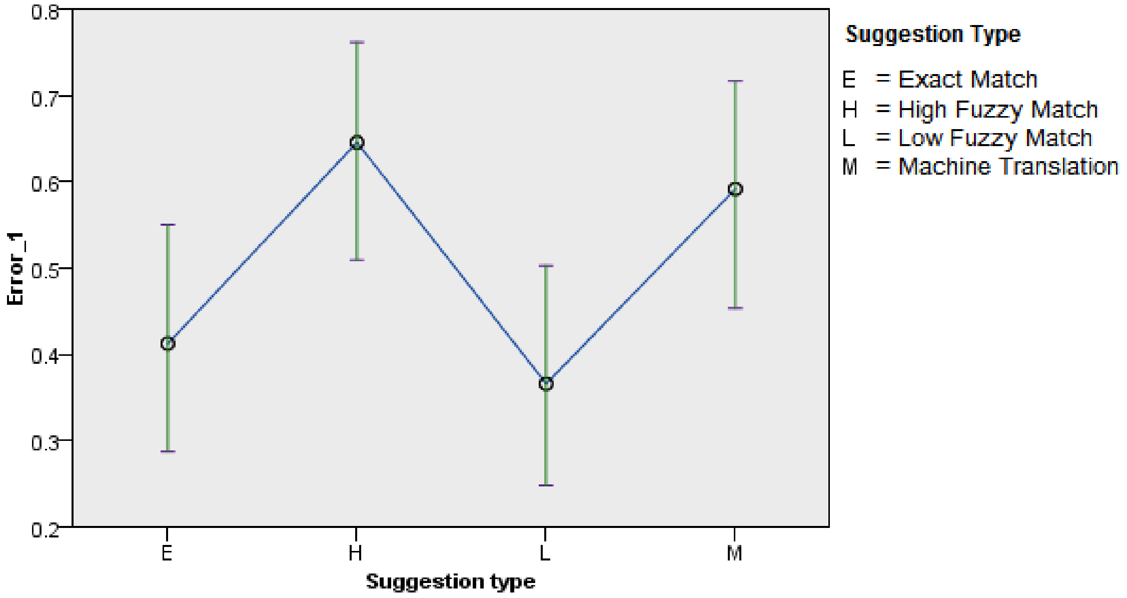


Table 52. Estimated marginal means for Error Score (binary), with Suggestion Type as a main effect

Suggestion Type	Mean	Std. Error
Exact Match (E)	0.412	0.069
High Fuzzy Match (H)	0.646	0.066
Low Fuzzy Match (L)	0.366	0.066
Machine Translation (M)	0.592	0.069

As indicated in Table 53, the mean differences are not statistically significant between Exact Matches and Low Fuzzy Matches, nor between High Fuzzy Matches and Machine Translation. In other words, the statistical model estimates that *the translators made errors less frequently when editing Exact Matches and Low Fuzzy Matches than they did when editing High Fuzzy Matches and Machine Translations*.

Table 53. Pairwise comparisons of estimated marginal means for Error Score (binary), with Suggestion Type as a main effect

Suggestion Types	Mean Difference	Mean Difference (%)	p*
E - H	*-0.233*	*-36%*	*0.001*
E - L	0.047	13%	0.746
E - M	*-0.179*	*-30%*	*0.011*
H - L	*0.280*	*77%*	*0.000*
H - M	0.054	9%	0.746
L - M	*-0.226*	*-38%*	*0.001*

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

5.4.7. Error score (non-zero cases)

Similarly to the strategy used for Typing Effort, the second step for analysing the non-normally distributed data for Error Score was to analyse the data points that had at least one error (using the log-transformed data) (see Figure 31). From a total of 560 data points, 277 were equal to zero and were eliminated and the remaining 283 constituted the new dataset. The resulting numeric variable was used as the dependent variable in a linear mixed-effects model, assuming a normal distribution. Following the same procedure described in the previous sections, the statistical model was used to test for main effects and then for interactions, and the non-significant effects were eliminated progressively. Only Suggestion Type was determined to be a significant effect (F = 5.159; p = 0.002), while no interaction effects were found to be statistically significant, similarly to what was found when Error Score was treated as a binary variable (section 5.4.6). The non-significant effects of Task and the interaction between Task and Suggestion Type are kept in the model all the same, as they are related to my sub-hypothesis 3a. The results of the final model are indicated in Table 54.

Table 54. Type III tests of fixed effects on Error Score (non-zero, log), including a significant main effect (final model)

Predictor	Numerator df	Denominator df	F	p*
Task	1	268	2.807	0.095
Suggestion Type	*3*	*266*	*5.159*	*0.002*
Task × Suggestion Type	3	267	0.569	0.636

* Rows in italics indicate significant results ($\alpha = 0.05$).

The significant main effect of Suggestion Type will be analysed in the following subsection.

5.4.7.1. Suggestion Type (main effect)

The effect of Suggestion Type on Error Score is indicated graphically in Figure 33 and through the estimated marginal means in Table 55. We should bear in mind that the original meaning of the variable has now been recovered (after undoing the logarithmic transformation), which is the number of errors per 100 words of source text. As an example of how to interpret the results, Table 55 indicates that translators made 5.1 errors per 100 words on average when translating Exact Match segments, considering only those segments where at least one error was made (this in turn corresponds to 41.2% of the total number of Exact Match segments, according to Table 52).

Figure 33. Estimated means for Error Score (non-zero, log), with Suggestion Type as a main effect

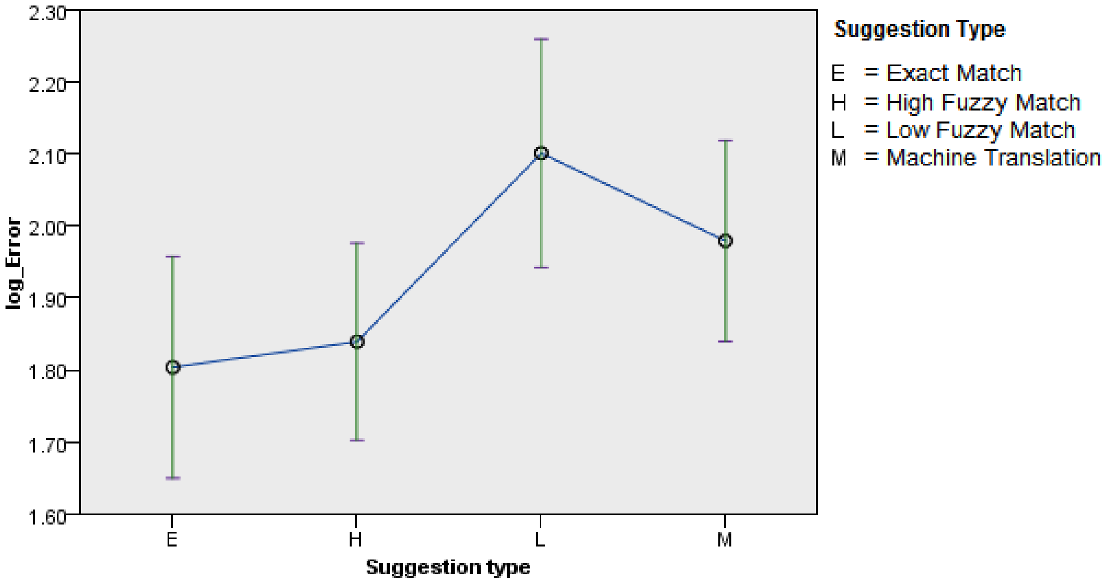


Table 55. Estimated marginal means for Error Score (non-zero), with Suggestion Type as a main effect

Suggestion Type	Mean (log)	Std. Error	Mean (errors / 100 words)
Exact Match (E)	1.804	.078	5.1
High Fuzzy Match (H)	1.839	.070	5.3
Low Fuzzy Match (L)	2.101	.081	7.2
Machine Translation (M)	1.979	.071	6.2

Table 56 indicates that the mean differences are statistically significant only between Exact Matches and Low Fuzzy Matches and between High Fuzzy Matches and Low Fuzzy Matches.

Table 56. Pairwise comparisons of estimated marginal means for Error Score (non-zero), with Suggestion Type as a main effect

Suggestion Types	Mean Difference	Mean Difference (%)	p*
E - H	-0.2	-4%	0.650
E - L	*-2.1*	*-29%*	*0.005*
E - M	-1.2	-19%	0.115
H - L	*-1.9*	*-26%*	*0.007*
H - M	-0.9	-15%	0.154
L - M	0.9	15%	0.283

* The mean difference is significant at the .05 level, with Bonferroni adjustments for multiple comparisons. Rows in italics indicate significant results.

The results presented in the current section combined with those presented in section 5.4.6, when Error Score was treated as a binary variable but all data points were used, can be interpreted as follows: High Fuzzy Matches and Machine Translations have the highest percentage of segments with errors (Figure 32), but when we look only at the segments that contain errors (Figure 33), High Fuzzy Matches have one of the smallest number of errors. On the other hand, Low Fuzzy Matches have the lowest percentage of segments with errors (Figure 32), but they have the highest number of errors when only the segments that contain errors are considered (Figure 33).

5.5. Additional information from eye tracking and screen recordings

In section 5.4 above, I present a comprehensive statistical analysis of the translators' performances under the two main test conditions (with and without metadata) and take into account other potential intervening factors. The purpose of the current section is to present a more narrative, humanised account of their performance in specific segments. I will analyse specific passages in the recordings with the help of the available eye-tracking data, but with no statistics based on the data for individual segments. The analysis is intended as an illustration of how eye tracking can help elucidate potential reasons behind some of the translators' behaviours, more than to provide an extensive account of all the available material generated in the experiment. For this analysis, I chose five segments that illustrate how the performances can be affected by the way the translators interact with the translation tool. In some cases, I will present a brief quantitative description of the translators' behaviours in the particular segment before proceeding to the qualitative analysis and the narrative explanation of what can be seen in the videos. The eye-tracking data will be used to help understand certain phenomena and will be illustrated by gaze plots.

The decision to analyse only certain segments in detail instead of presenting an extensive analysis of all the eye-tracking and video material available was made for several reasons. The first is of practical nature, as it would have required an excessive amount of time to analyse in detail all segments in all the recordings. “Data explosion” is a phrase normally used to describe such an excess of data in process research. An alternative would have been to do a global analysis in a general quantitative way (for example, by looking at the fixations per area of interest to see how the participants used translation metadata when available), according to the type of suggestion both in the presence and absence of metadata. However, this would still have required that all segments (28 segments per task per participant, i.e. 560 segments in total) be manually identified and divided in Tobii Studio. Another major reason for choosing to analyse specific segments and participants instead of the whole recordings is that some of my eye-tracking recordings had problems of data loss and inconsistent calibration, which would have introduced numerous errors in a quantitative analysis. These problems are illustrated in the current section, and the reasons why they occurred were discussed in section 4.6.3.

5.5.1. Example 1

The first segment to be analysed corresponds to a high-percentage fuzzy match of 95% (as calculated by IBM TranslationManager). It contains 18 source words and is the 12th segment of SourceText42. The text presentation in the tool is as follows:

Source text in the translation memory:

In the <span class=""keyword"">Tivoli® Enterprise Portal</span>, click the Navigator item of the monitoring agent and click Start or Restart

Source text in the active segment:

In the <span class=""keyword"">Tivoli® Enterprise Portal</span>, **right-**click the Navigator item of the monitoring agent and click Start or Restart

Translation suggestion:

```
En el <span class=""keyword"">Tivoli Enterprise Portal</span>,
pulse el elemento de Navigator del agente de supervisión y pulse
Iniciar o Reiniciar
```

The only difference between the original source text and the source text in the active segment is the verb "right-click" instead of just "click", as indicated in bold above. It is also worth noting that there is a missing final stop in the English original, and the suggested translation does not contain the registered trademark symbol (®), as this is IBM's general guideline for the translation of the Tivoli family of products into Spanish. The idea now is to analyse how the translators handled this segment when metadata were available in comparison to when metadata were not available.

Participants P01, P03, P04, P07 and P08 translated this text in the Visual task (with metadata), while P02, P05, P06, P09 and P10 translated it in the Blind task (no metadata). Table 57 presents the quantitative information comparing the performances between the two tasks. Table 58 provides the final translations produced by each translator.

Table 57. Detailed view of translators' performances when translating a High Fuzzy Match under both conditions (Example 1)

VISUAL

Participant	Visits	Characters in target	Time (s)	Keystrokes	Errors
P01	1	173	30	32	0
P03	1	177	38	126	2
P04	2	163	25	22	1
P07	1	176	16	31	1.5
P08	3	179	101	81	0

BLIND

Participant	Visits	Characters in target	Time (s)	Keystrokes	Errors
P02	2	171	66	38	0
P05	2	170	33	33	2
P06	1	173	31	32	2
P09	1	142	16	1	4
P10	2	142	30	21	4

"Visits" refers to how many times the translators activated the segment. One visit means that the translator activated it once, produced their translation and did not activate the segment again later; two or more visits means that the translator activated the segment more than once, usually in the proof-reading or self-revising phase at the end.[19] "Time" is counted in seconds from the moment the segment is "opened" (activated) to the moment it is "closed" (deactivated and saved into the translation memory). "Keystrokes" refers to how many keyboard keys were pressed (except for control keys – see section 4.8.1), so this includes deletions as well as additions. "Errors" indicates the mean of the error scores obtained from the evaluations by both reviewers (see section 4.8.2).

Table 58. Final translations based on a High Fuzzy Match under both conditions (Example 1)

VISUAL

Participant	Final translation
P01	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse con el botón derecho del ratón el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P03	En el <span class="keyword">Tivoli® Enterprise Portal</span>, pulse con el botón derecho del ratón el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P04	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse con el botón derecho el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P07	En el <span class="keyword">Tivoli Enterprise Portal</span>, pulse con el botón derecho del ratón el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P08	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse con el botón derecho del ratón sobre el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar

BLIND

Participant	Final translation
P02	En <span class="keyword">Tivoli® Enterprise Portal</span>, pulse con el botón derecho sobre el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P05	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse con el botón derecho del ratón el elemento Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P06	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse con el botón derecho del ratón el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P09	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar
P10	En <span class="keyword">Tivoli Enterprise Portal</span>, pulse el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar

The data presented in Table 57 is summarised in Table 59, which indicates the extreme values, the mean and the median, across the five participants who translated the segment under scrutiny in each of the tasks. While the means indicate clear differences between the two tasks, we should keep in mind that the mean is very sensitive to extreme

[19] Translators can also change their translation in a mode called "post-editing", without activating the segments.

values (consequently, in our case, to personal differences between the participants). Because the median is less sensitive to extreme values, it gives a more neutral (individual-independent) view of the differences between the two tasks, and in Table 59 they indeed indicate much slighter differences between the two conditions.

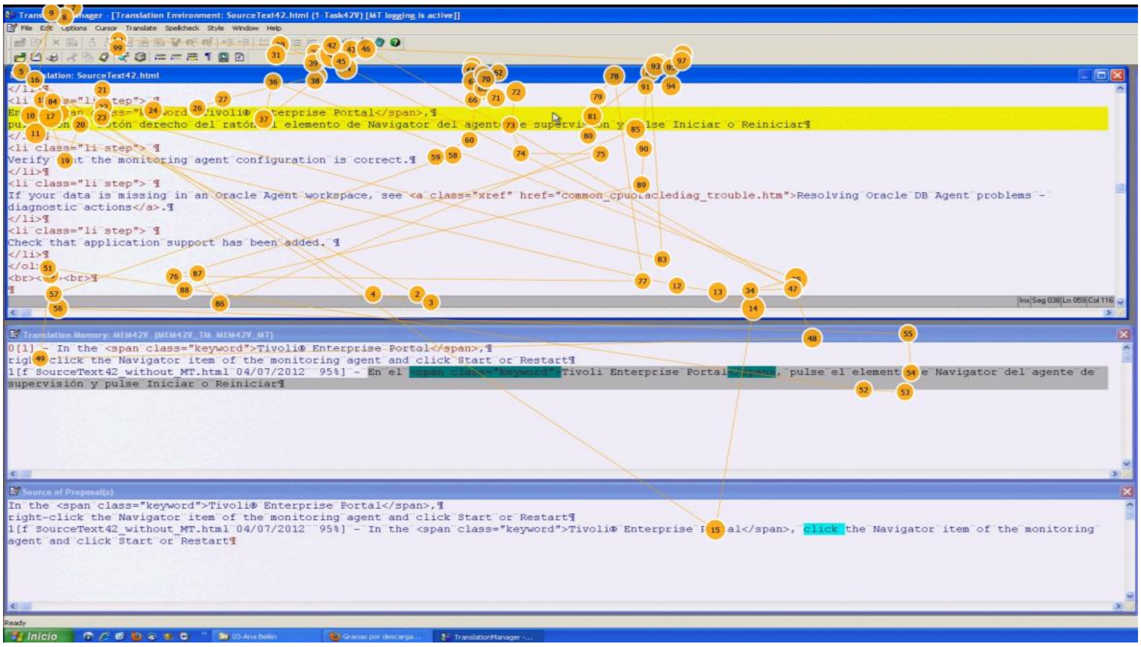
Table 59. Summarised view of translators' performances when translating a High Fuzzy Match under both conditions (Example 1)

	Time (s)			Keystrokes			Errors		
	Range	Mean	Median	Range	Mean	Median	Range	Mean	Median
Visual	16 - 101	42	30	22 - 126	58	32	0 - 2	0.9	1
Blind	16 - 66	35	31	1 - 38	25	32	0 - 4	2.4	2

I will start by comparing the performances of P03 and P09, prompted by the observation that P03 pressed 126 keys to produce her final translation while P09 pressed just one key to produce hers. The explanation can be found by watching the video recordings of their performances with the help of the eye-tracking data. The gaze plots of their performances are presented in Figure 34 and Figure 35 below.

P03 translated this segment in the Visual task and she had the particularity of not inserting the translation suggestion into the editing area; instead, she typed all her translation on top of the source text, which comes by default when the segment is opened. The reason why she still pressed fewer keys (126) than the number of characters needed to produce her translation (177) is that she took advantage of some of the existing English text, such as the tags, the name of the product ("Tivoli Enterprise Portal") and some initial letters of the English words when they were identical to what she wanted to write in Spanish. P09, on the other hand, was working on the Blind task and had the translation suggestion already inserted in the segment. She first read the suggested target text for four seconds, then double-clicked the article "el" before the product name with the mouse and pressed the Delete key (only one key press in all). Then she read the target text for two more seconds and moved her eyes to the source text (the middle pane in Figure 35), where she spent 2.5 seconds, then she alternated between the target and the source until she moved on to the next segment. P03, on the other hand, had the source text within the segment where she was producing the target; yet she was regularly looking at the middle pane while she typed, so in her case we might suppose she was consulting the suggestion rather than checking the source text. (Despite some shift in the gaze data, we assume that the fixations shown over blank areas in the centre of the screen actually correspond to the Translation Memory pane, a little further below.)

Figure 34. Gaze plot of P03 while translating a High Fuzzy Match in the Visual task (Example 1)

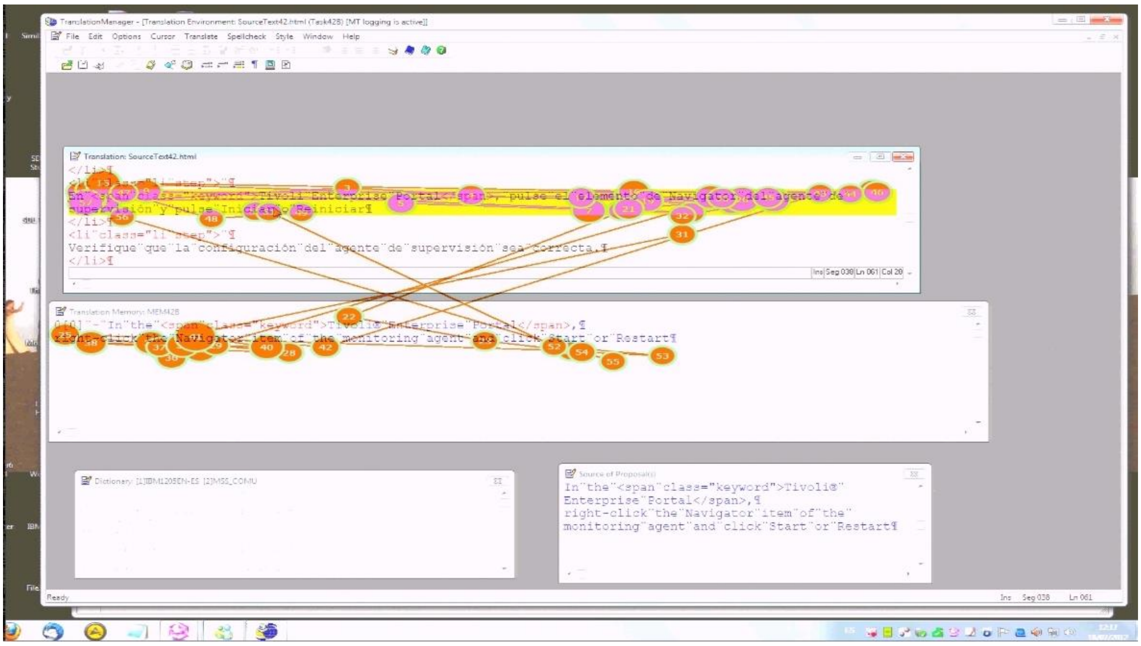

TranslationManager - [Translation Environment: SourceText42.html (1-Task42V) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText42.html
Ins Seg 038 Ln 059 Col 116
Translation Memory: MEM42V (MEM42V_TM MEM42V_MT)
Source of Proposal(s)
Ready
Inicio
03-Ana Belén
Gracias por descarga...
TranslationManager -...

```
</li>¶
<li class="li step"> ¶
En <span class="keyword">Tivoli® Enterprise Portal</span>,¶
pulse con el botón derecho del ratón el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar¶
</li>¶
<li class="li step"> ¶
Verify that the monitoring agent configuration is correct.¶
</li>¶
<li class="li step"> ¶
If your data is missing in an Oracle Agent workspace, see <a class="xref" href="common_cpuoracleiag_trouble.htm">Resolving Oracle DB Agent problems -
diagnostic actions</a>.¶
</li>¶
<li class="li step"> ¶
Check that application support has been added. ¶
</li>¶
</ol>¶
<br><br><br>¶
¶
```

```
0[1] - In the <span class="keyword">Tivoli® Enterprise Portal</span>,¶
right-click the Navigator item of the monitoring agent and click Start or Restart¶
1[f SourceText42_without_MT.html 04/07/2012  95%] - En el <span class="keyword">Tivoli Enterprise Portal</span>, pulse el elemento de Navigator del agente de
supervisión y pulse Iniciar o Reiniciar¶
```

```
In the <span class="keyword">Tivoli® Enterprise Portal</span>,¶
right-click the Navigator item of the monitoring agent and click Start or Restart¶
1[f SourceText42_without_MT.html 04/07/2012  95%] - In the <span class="keyword">Tivoli® Enterprise Portal</span>, click the Navigator item of the monitoring
agent and click Start or Restart¶
```

Figure 35. Gaze plot of P09 while translating a High Fuzzy Match in the Blind task (Example 1)


TranslationManager - [Translation Environment: SourceText42.html (Task42B) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText42.html
Ins Seg 038 Ln 061 Col 20
Translation Memory: MEM42B
Dictionary: [1]IBM1205EN-ES [2]MSS_COMU
Source of Proposal(s)
Ready
Ins Seg 038 Ln 061

```
</li>¶
<li class="li step"> ¶
En <span class="keyword">Tivoli Enterprise Portal</span>, pulse el elemento de Navigator del agente de
supervisión y pulse Iniciar o Reiniciar¶
</li>¶
<li class="li step"> ¶
Verifique que la configuración del agente de supervisión sea correcta.¶
</li>¶
```

```
0[0] - In the <span class="keyword">Tivoli® Enterprise Portal</span>,¶
right-click the Navigator item of the monitoring agent and click Start or Restart¶
```

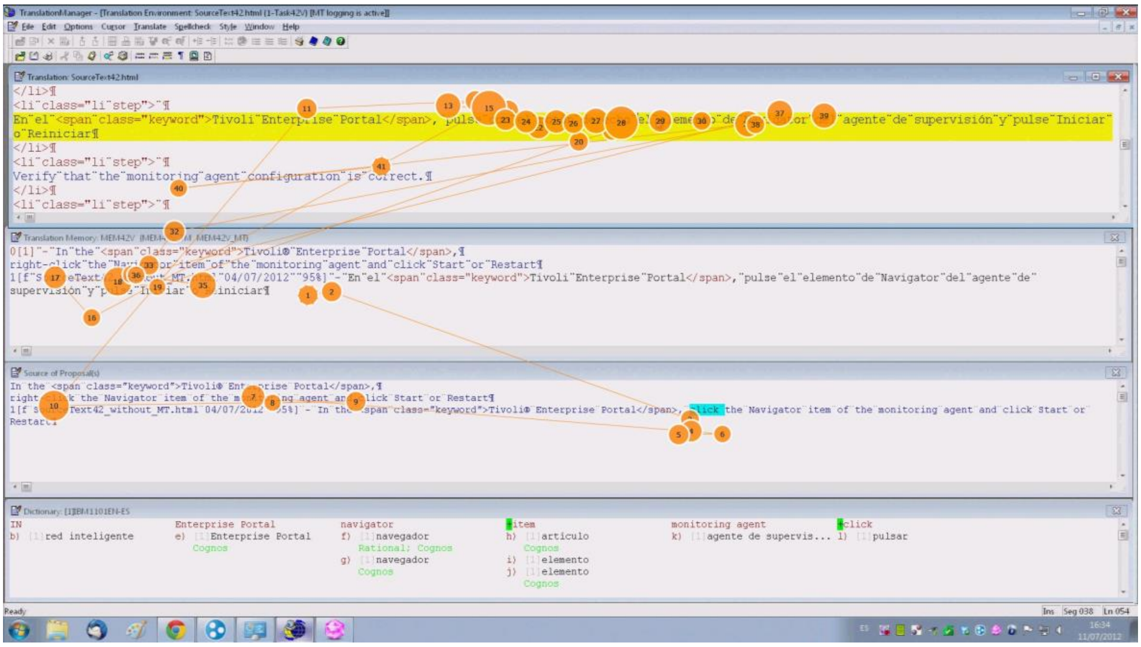
```
In the <span class="keyword">Tivoli®
Enterprise Portal</span>,¶
right-click the Navigator item of the
monitoring agent and click Start or Restart¶
```

The gaze plot for P09, presented in Figure 35, also illustrates another eye-tracking issue, where the full width of the screen was not captured in some cases (see section 4.6.3). The figure shows that the screen image is cut on the left, as can be seen by the missing Start button. The full window of the translation tool was still captured because I asked the translator to move the window to the right, within the captured area, which I could monitor dynamically before starting the recording. However, the eye-tracking data seem to have "shrunk" towards the right, as the gaze plot shows no fixations on the initial parts of the sentences. It is very unlikely that the translator had a gaze pattern that skipped the left-hand part of the window systematically. Because of this issue, the eye-tracking data do not allow us to know exactly which words were being fixated, but the gaze plot is still useful to indicate the distribution of visual attention between the two panes.

Having explained the huge difference in the number of keystrokes between the two translators, let us now have a look at why their final translations were so different in length (177 for P03 vs. 142 for P09). The explanation is that P09, working without metadata, did not spot the difference between the two source texts: she kept the translation for "click" ("pulse") instead of repairing it to correspond to the new original "right-click". P03, who was working with metadata, correctly changed the available suggestion to "pulse con el botón derecho del ratón" (the recommended translation for "right-click" in Spanish). The remaining difference in the character count is due to the article "el" before the product name and the registered trademark symbol, both of which P03 did not delete. P10, who was also working in the Blind task, made the same error as P09, while none of the translators working in the Visual task made this error, which indicates that metadata might have played an important role in this segment.

Another translator, P04 has the typical behaviour one would expect from a translator working in a translation memory system with metadata. His gaze plot is presented in Figure 36. It shows that the translator consults the suggestion, spots the difference between the source texts and changes only the word that was different.

Figure 36. Gaze plot of P04 while translating a High Fuzzy Match in the Visual task (Example 1)


TranslationManager - [Translation Environment: SourceText42.html (1-Task42V) [MT logging is active]]
File
Edit
Options
Cursor
Translate
Spellcheck
Style
Window
Help
Translation: SourceText42.html
Translation Memory: MEM42V (MEM4 M_MEM42V_MT)
Source of Proposal(s)
Dictionary: [1]IBM1101EN-ES
Ready
Ins
Seg 038
Ln 054
ES
16:34
11/07/2012
11
13
15
23
24
25
26
27
28
29
30
37
38
39
20
41
40
32
33
17
18
36
19
35
16
1
2
7
8
9
10
5
6

```
</li>¶
<li class="li step"> ¶
En el <span class="keyword">Tivoli Enterp ise Portal</span>, puls  e  eme  de  or agente de supervisión y pulse Iniciar o Reiniciar¶
</li>¶
<li class="li step"> ¶
Verify that the monitoring agent configuration is co rect.¶
</li>¶
<li class="li step"> ¶
```

```
0[1] - In the <span class="keyword">Tivoli® Enterprise Portal</span>,¶
right-click the Navi or item of the monitoring agent and click Start or Restart¶
1[f S eText  MT.  "04/07/2012" "95%] - En el <span class="keyword">Tivoli Enterprise Portal</span>, pulse el elemento de Navigator del agente de supervisión y p  In iar  iniciar¶
```

```
In the <span class="keyword">Tivoli® Ent rise Portal</span>,¶
right  k the Navigator item of the m  ng agent an  lick Start or Restart¶
1[f s  Text42_without MT.html 04/07/20  95%] - In the span class="keyword">Tivoli® Enterprise Portal</span>, lick the Navigator item of the monitoring agent and click Start or Restart
```

```
IN                          Enterprise Portal          navigator                item                        monitoring agent          click
b)  [1] red inteligente     e)  [1] Enterprise Portal  f)  [1] navegador        h)  [1] artículo            k)  [1] agente de supervis... l)  [1] pulsar
                                Cognos                         Rational; Cognos         Cognos
                                                       g)  [1] navegador        i)  [1] elemento
                                                               Cognos           j)  [1] elemento
                                                                                        Cognos
```

5.5.2. Example 2

Now let us look at how the same translator P04 translated a similar segment in the Blind task. For this comparison I will take segment 24 of SourceText31, which is also a High Fuzzy Match (97%), with the following characteristics:

Source text in the translation memory:

`An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application written in a high-level language to use specific data or functions of the operating system or the licensed program.`

Source text in the active segment:

`An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application` **`program`** `written in a high-level language to use specific data or functions of the operating system or the licensed program.`

Translation suggestion:

`Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que una aplicación escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.`

The only difference between the source text in the translation memory and the source text in the active segment in this case is the addition of the word “program”, which creates the term “application program” instead of just “application”, as indicated in bold above. Table 60 presents the quantitative information on the participants’ performances while translating this segment. Table 61 lists the final translations produced by each translator.

Table 60. Detailed view of translators' performances when translating a High Fuzzy Match under both conditions (Example 2)

VISUAL

Participant	Visits	Characters in target	Time (s)	Keystrokes	Errors
P02	2	274	47	19	0
P05	1	274	25	15	0
P06	1	274	41	31	0
P09	1	274	22	15	0
P10	1	272	28	14	0.5

BLIND

Participant	Visits	Characters in target	Time (s)	Keystrokes	Errors
P01	1	263	19	0	0
P03	1	274	55	22	1
P04	1	268	106	25	0.5
P07	1	263	53	0	0
P08	1	263	118	39	0

Contrary to what happened when P04 translated a similar High Fuzzy Match segment in the Visual task, here in the Blind task he seems to have overlooked the difference between the two source texts. It is true that there is no real difference in meaning between “application” and “application program” and this might be the reason why the translator did not bother to change the suggested translation. However, it is an unlikely coincidence that all participants who translated this segment in the Blind task (P01, P04, P07 and P08) but one (P03) decided to leave “application program” translated as “aplicación”, while all participants who translated the segment in the Visual task (P02, P05, P06 and P09) but one (P10) preferred to change it to “programa de aplicación”. It is much more reasonable to assume that the highlighted difference between the source texts in the Visual task accounts for the difference in behaviour.

Table 61. Final translations based on a High Fuzzy Match under both conditions (Example 2)

VISUAL

Participant	Final translation
P02	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que un **programa de aplicación** escrito en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P05	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que un **programa de aplicación** escrito en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P06	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que un **programa de aplicación** escrito en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P09	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que un **programa de aplicación** escrito en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P10	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia separada que permite que una **aplicación** escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.

BLIND

Participant	Final translation
P01	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que una **aplicación** escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P03	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que un **programa de aplicación** grabado en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P04	Una API es una interfaz funcional suministrada por el sistema operativo u otro programa bajo licencia diferente que permite a una **aplicación** escrita en un lenguaje de alto nivel utilizar datos o funciones específicos del sistema operativo o del programa bajo licencia.
P07	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que una **aplicación** escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
P08	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que una **aplicación** escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.

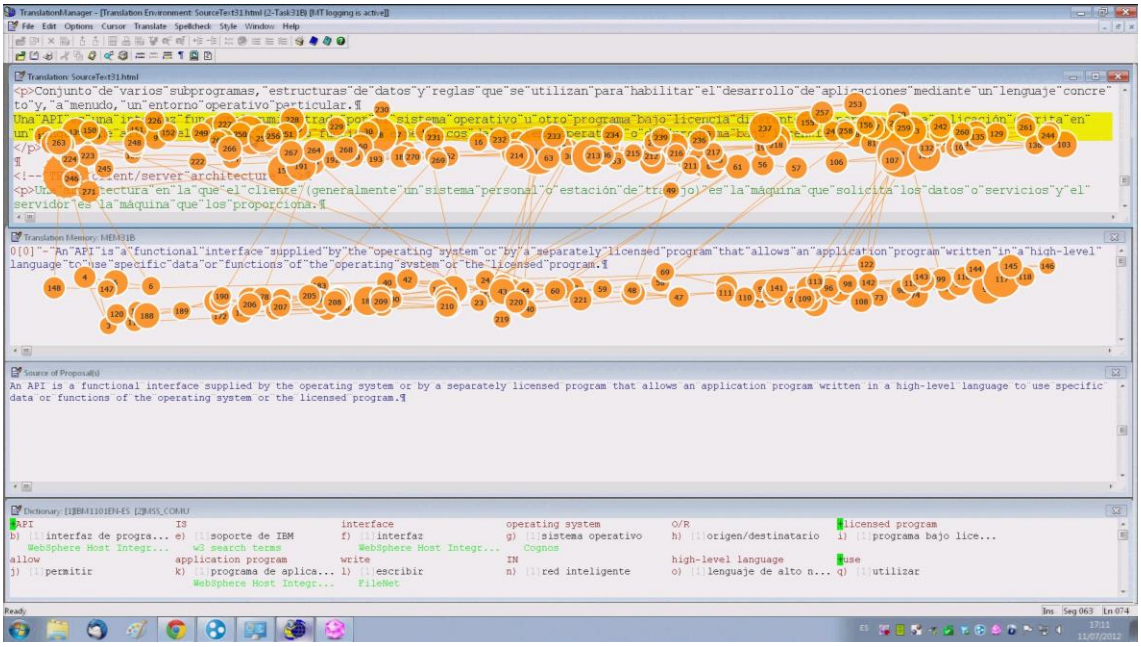
The gaze plot of P04’s performance while translating the segment is presented in Figure 37. It shows that his visual attention was distributed much more homogeneously between the target text (upper pane) and the source text (upper middle pane) when compared to the gaze plot presented in Figure 36, which shows the same translator working on the Visual task. Here in the Blind task he does not even look at the lower middle pane, which in the Visual task contained translation metadata, including the differences between the source texts. The lack of metadata seems to have been responsible for the more frequent gaze switches between the source and the target texts and for the translator’s failure to identify a minor change in the source text. On a side note, it is interesting to mention that P04 did not check the client's glossary, which was open in the lower pane and contained, among other terms, the recommended translation for "application program".

As for the quality of the eye-tracking data, Figure 37 corresponds to the second task performed by this translator and shows less precise data than Figure 36, which corresponds to the first task. As a general rule, the eye-tracker calibration became worse with time and the data presented some "drift", a problem reported in other studies (Hvelplund 2014: 210–211). This is acknowledged by the tool manufacturer, who define it as "the gradual decrease in accuracy of the eye tracking data compared to the true eye position" (Tobii Technology 2010: 10–11). Drift is attributed to "variations in eye physiology (e.g. degree of wetness, tears) and variations in the environment (e.g. sunlight variations)" (loc. cit.) and tends to become worse with longer recording times. To minimise the problem, the recommendation is to calibrate frequently, but this is not always a possible solution, especially when one cannot interrupt a translation session.

Table 61 above shows that only one translator (P10) of those working in the Visual task did not change the original translation suggested by the TM, "aplicación", to "programa de aplicación", which would be the natural choice prompted by the highlighted difference between the source texts (and also recommended by the glossary). Two reasons might explain his behaviour. The first reason is suggested by his screen recording, as it shows that he was concerned about changing the translation for the word "separately" and probably paid less attention to the rest of the sentence. A second reason can be inferred from his eye movement behaviour, illustrated by the gaze plot in Figure 38. It suggests that the translator did not consult the metadata in the lower pane, which indicated the difference between the two source texts with a question mark highlighted in magenta.

The gaze plot in Figure 38 suffers from the same issue discussed in the previous section, where the image from the translator's screen was not captured in its entirety by the eye-tracking application. In this case, the image is cut on the right side of the translator's screen (notice that the clock and typical icons are not there at the bottom right) and the eye-tracking data is shifted towards the left, although the displacement is not as strong as the example shown in Figure 35. In this case, despite the low accuracy of the eye-tracking data, it is safe to assume that, even if the translator fixated at the "f" at the beginning of the line in the middle pane (which indicates a fuzzy match), he did not pay attention to the difference between the two source texts in the lower pane.

Figure 37. Gaze plot of P04 while translating a High Fuzzy Match in the Blind task (Example 2)


TranslationManager - [Translation Environment: SourceText31.html (2-Task31B) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText31.html
Translation Memory: MEM31B
Source of Proposal(s)
Dictionary: [1]IBM1101EN-ES [2]MSS_COMU
Ready
Ins Seg 063 Ln 074
17:11
11/07/2012

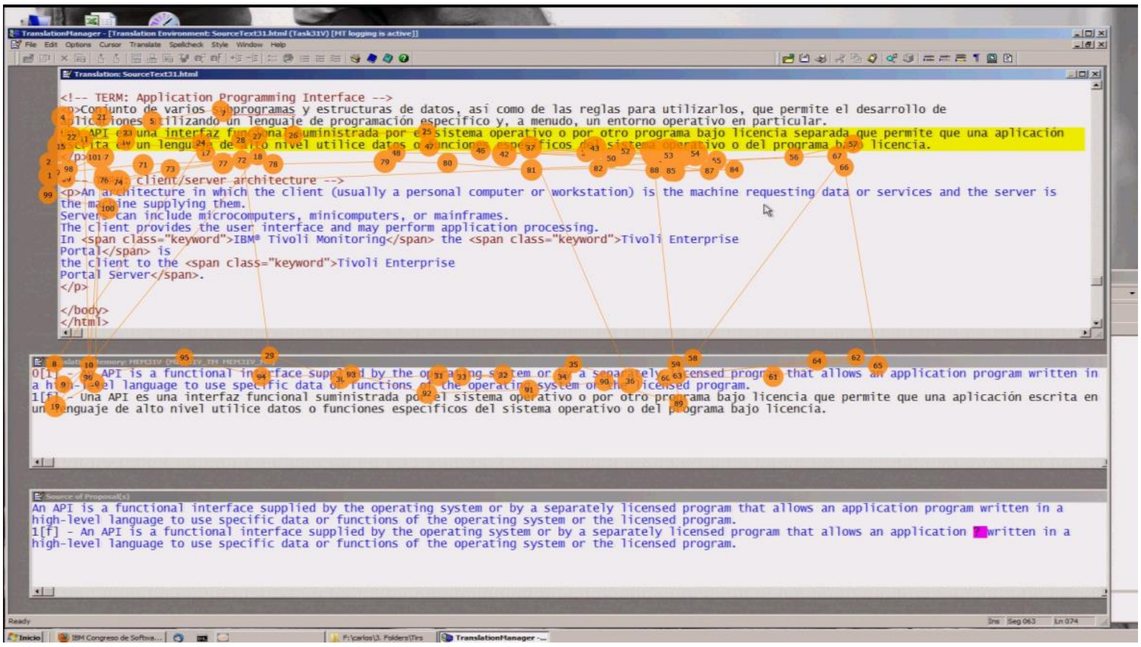
<p>Conjunto de varios subprogramas, estructuras de datos y reglas que se utilizan para habilitar el desarrollo de aplicaciones mediante un lenguaje concreto y, a menudo, un entorno operativo particular.¶
Una API [illegible] una interfaz fun[illegible] trad[illegible] por [illegible] sistema operativo u otro programa bajo licencia di[illegible] licación [illegible]ita en
un [illegible] el [illegible] cos [illegible] operat[illegible] o [illegible] ma ba[illegible]
</p>
¶
<!-- [illegible] client/server architectur[illegible]
<p>Un[illegible]tectura en la que el cliente (generalmente un sistema personal o estación de tra[illegible]jo) es la máquina que solicita los datos o servicios y el servidor es la máquina que los proporciona.¶

0[0] - An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.¶

An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.¶

API
b) [1] interfaz de progra...
WebSphere Host Integr...
IS
e) [1] soporte de IBM
w3 search terms
interface
f) [1] interfaz
WebSphere Host Integr...
operating system
g) [1] sistema operativo
Cognos
O/R
h) [1] origen/destinatario
licensed program
i) [1] programa bajo lice...
allow
j) [1] permitir
application program
k) [1] programa de aplica...
WebSphere Host Integr...
write
l) [1] escribir
FileNet
IN
n) [1] red inteligente
high-level language
o) [1] lenguaje de alto n...
use
q) [1] utilizar

Figure 38. Gaze plot of P10 while translating a High Fuzzy Match in the Visual task (Example 2)


TranslationManager - [Translation Environment: SourceText31.html (Task31V) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText31.html
Source of Proposal(s)
Ready
Ins Seg 063 Ln 074
Inicio
IBM Congreso de Softwa...
F:\carlos\3. Folders\Trs
TranslationManager -...

<!-- TERM: Application Programming Interface -->
<p>Conjunto de varios subprogramas y estructuras de datos, así como de las reglas para utilizarlos, que permite el desarrollo de aplicaciones utilizando un lenguaje de programación específico y, a menudo, un entorno operativo en particular.
[illegible] API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia separada que permite que una aplicación escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.
</p>
<!-- [illegible] client/server architecture -->
<p>An architecture in which the client (usually a personal computer or workstation) is the machine requesting data or services and the server is the machine supplying them.
Servers can include microcomputers, minicomputers, or mainframes.
The client provides the user interface and may perform application processing.
In <span class="keyword">IBM® Tivoli Monitoring</span> the <span class="keyword">Tivoli Enterprise Portal</span> is
the client to the <span class="keyword">Tivoli Enterprise Portal Server</span>.
</p>

</body>
</html>

0[1] - An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.
1[f] - Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que una aplicación escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.

An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.
1[f] - An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application written in a high-level language to use specific data or functions of the operating system or the licensed program.

5.5.3. *Example 3*

The third example to be investigated corresponds to a machine-translation suggestion and is the 27th segment in SourceText31. It contains the following:

Source text in the active segment:

```
The client provides the user interface and may perform application
processing.
```

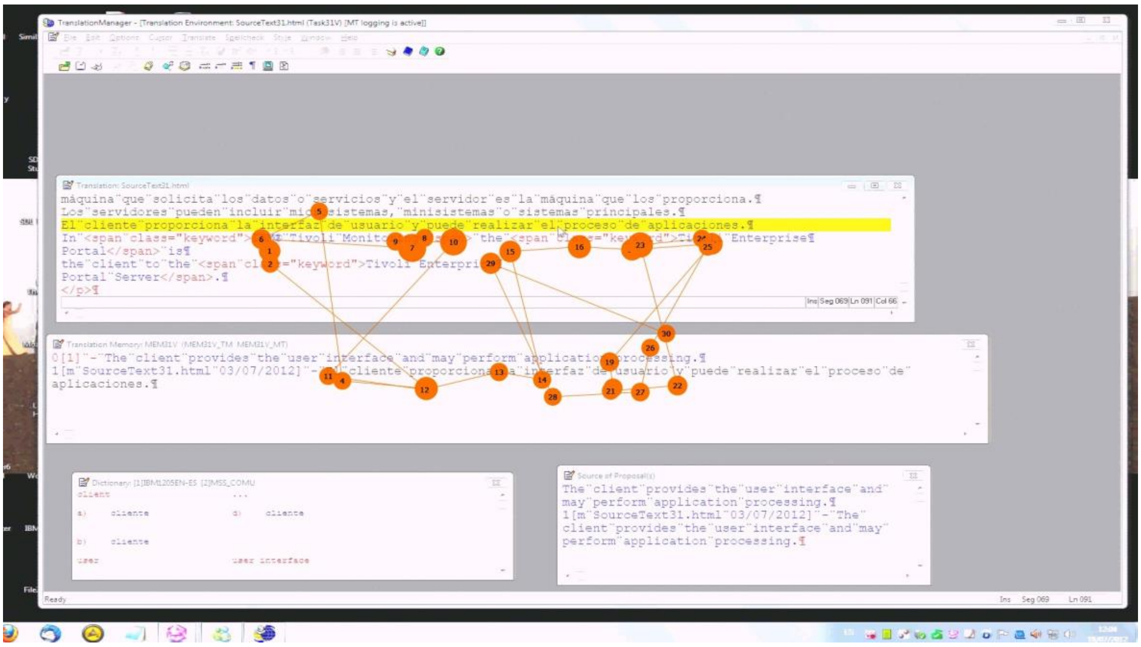
Translation suggestion:

```
El cliente proporciona la interfaz de usuario y puede realizar el
proceso de las aplicaciones.
```

In this case, there is no reference source text in the translation memory, because the suggestion was generated by machine translation. This particular segment is presented as an example because the translation suggestion produced by the machine translation engine happened to be exactly the same as the translation that was taken from the published manual in Spanish. In a real-world scenario, this suggestion would have been an exact match, but in the experiment it was flagged as a machine translation (for those seeing it in the Visual environment). I am especially interested in checking the performance of P09 when translating this segment in the Visual task, because she usually accepted exact matches extremely quickly, mainly because of the key combination she used to move around the text (see section 6.2.5). In the current segment, she did not touch the suggested translation, but she spent 10 seconds reading it, which corresponds to 90 seconds per 100 words, against an average of 9 seconds per 100 words for exact matches (her general average for machine translation suggestions was 174 seconds per 100 words). This behaviour suggests that the time spent on a segment does not depend only on the intrinsic characteristics of the translation suggestion, but also on the *trust* attributed to the type of suggestion.

The low accuracy of the eye-tracking data for this participant (see section 5.5.1) does not allow us to confirm whether she consulted the metadata, which in this case consists of the "m" at the beginning of the line in the middle pane. Unexpectedly, however, even if we were to "stretch" the gaze data back towards the left, she does not seem to have fixated where the "m" is. Therefore, in this particular case, it is not safe to affirm that the metadata have influenced the translator in her decision of how to handle the suggestion.

Figure 39. Gaze plot of P09 while translating a Machine Translation suggestion in the Visual task (Example 3)



TranslationManager - [Translation Environment: SourceText31.html (Task31V) [MT logging is active]]

Translation: SourceText31.html

```
máquina¨que¨solicita¨los¨datos¨o¨servicios¨y¨el¨servidor¨es¨la¨máquina¨que¨los¨proporciona.¶
Los¨servidores¨pueden¨incluir¨mi sistemas,¨minisistemas¨o¨sistemas¨principales.¶
El¨cliente¨proporciona¨la¨interfaz¨de¨usuario¨y¨puede¨realizar¨el¨proceso¨de¨aplicaciones.¶
In¨<span¨class="keyword">   ¨Tivoli¨Monito          >¨the¨<span¨cl  s="key   rd">Ti     ¨Enterprise¶
Portal</span>¨is¶
the¨client¨to¨the¨<span¨cl  ="keyword">Tivoli Enterpri
Portal¨Server</span>.¶
</p>¶
```

Ins Seg 069 Ln 091 Col 66

Translation Memory: MEM31V (MEM31V_TM MEM31V_MT)

```
0[1]¨-¨The¨client¨provides¨the¨user¨interface¨and¨may¨perform¨applicatio  rocessing.¶
1[m¨SourceText31.html¨03/07/2012]¨-   ¨cliente¨proporciona  a¨interfaz¨de¨usuario¨y¨puede¨realizar¨el¨proceso¨de¨
aplicaciones.¶
```

Dictionary: [1]IBM1205EN-ES [2]MSS_COMU

```
client                          ...
a)    cliente               d)    cliente
b)    cliente
user                            user interface
```

Source of Proposal(s)

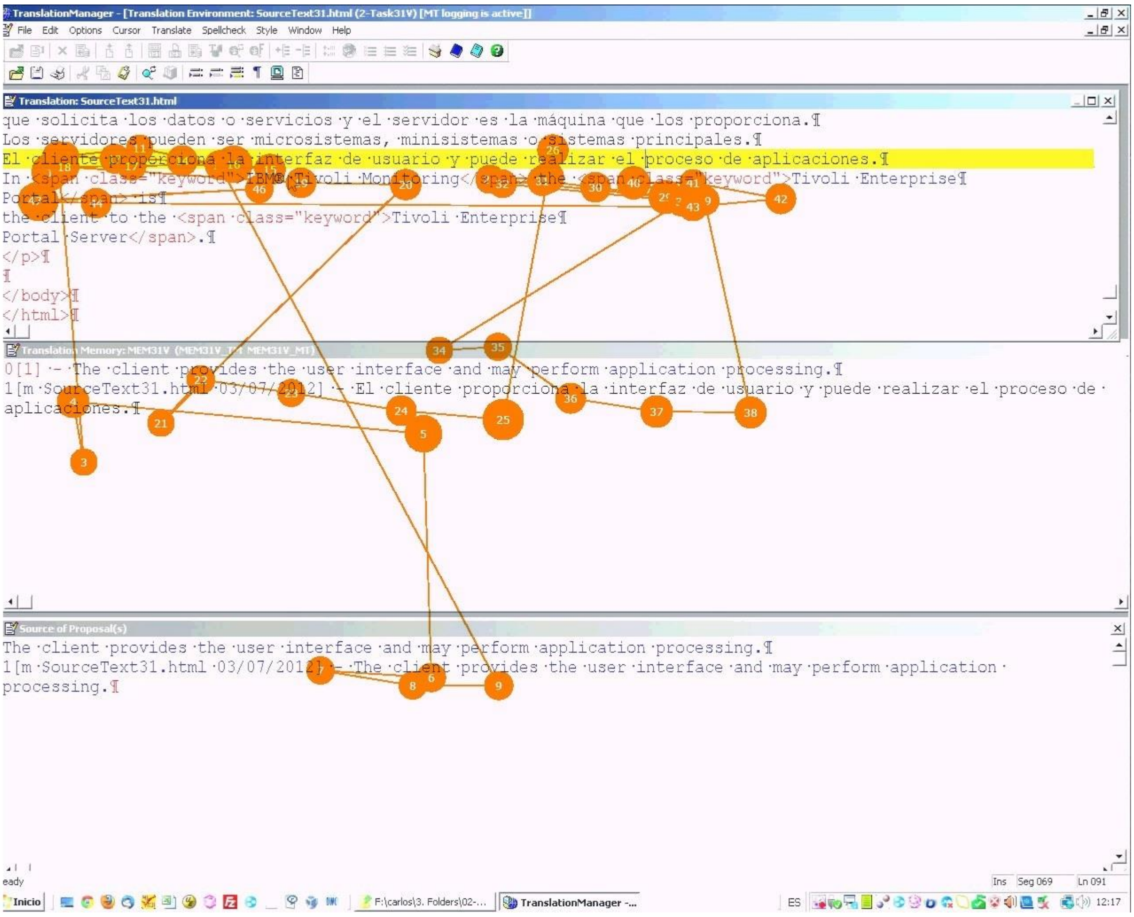
```
The¨client¨provides¨the¨user¨interface¨and¨
may¨perform¨application¨processing.¶
1[m¨SourceText31.html¨03/07/2012]¨-¨The¨
client¨provides¨the¨user¨interface¨and¨may¨
perform¨application¨processing.¶
```

Ready

Ins Seg 069 Ln 091

Let us look at another translator whose eye-tracking recording was more precise. Figure 40 shows the gaze plot of P02 translating the same segment. It indicates that, in an initial phase, the translator had a quick glance at the active segment in the upper pane (fixations 1-2), then moved to the middle pane (fixations 3-5), then to the lower pane (fixations 6-9), then back to the upper pane. It seems that he was initially concerned with the phrase “The client provides”, because this is the only chunk he checks in the lower pane and because of the high number of fixations on this zone in the upper pane. The other part of the segment that called his attention was around “puede realizar” (“may perform”) and then “proceso de aplicaciones” (“application processing”). He actually puts the cursor before the “p” in “proceso” and thinks for a while about whether he should add something or change the word. He adds a space then deletes it, before moving to the next segment. It is easier to visualise all this activity by inspecting his gaze behaviour dynamically in the recordings, where it is possible to choose the time interval for the cumulative fixations that are displayed, thus allowing less superposition than what is seen in Figure 40.

Figure 40. Gaze plot of P02 while translating a Machine Translation suggestion in the Visual task (Example 3)


TranslationManager - [Translation Environment: SourceText31.html (2-Task31V) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText31.html
Translation Memory: MEM31V (MEM31V_TM MEM31V_MT)
Source of Proposal(s)
eady
Ins Seg 069 Ln 091
Inicio
F:\carlos\3. Folders\02-...
TranslationManager -...
ES
12:17
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

```
que·solicita·los·datos·o·servicios·y·el·servidor·es·la·máquina·que·los·proporciona.¶
Los·servidores·pueden·ser·microsistemas,·minisistemas·o·sistemas·principales.¶
El·cliente·proporciona·la·interfaz·de·usuario·y·puede·realizar·el·proceso·de·aplicaciones.¶
In·<span·class="keyword">IBM®·Tivoli·Monitoring</span>·the·<span·class="keyword">Tivoli·Enterprise¶
Portal</span>·is¶
the·client·to·the·<span·class="keyword">Tivoli·Enterprise¶
Portal·Server</span>.¶
</p>¶
¶
</body>¶
</html>¶
```

```
0[1]·-·The·client·provides·the·user·interface·and·may·perform·application·processing.¶
1[m·SourceText31.html·03/07/2012]·-·El·cliente·proporciona·la·interfaz·de·usuario·y·puede·realizar·el·proceso·de·
aplicaciones.¶
```

```
The·client·provides·the·user·interface·and·may·perform·application·processing.¶
1[m·SourceText31.html·03/07/2012]·-·The·client·provides·the·user·interface·and·may·perform·application·
processing.¶
```

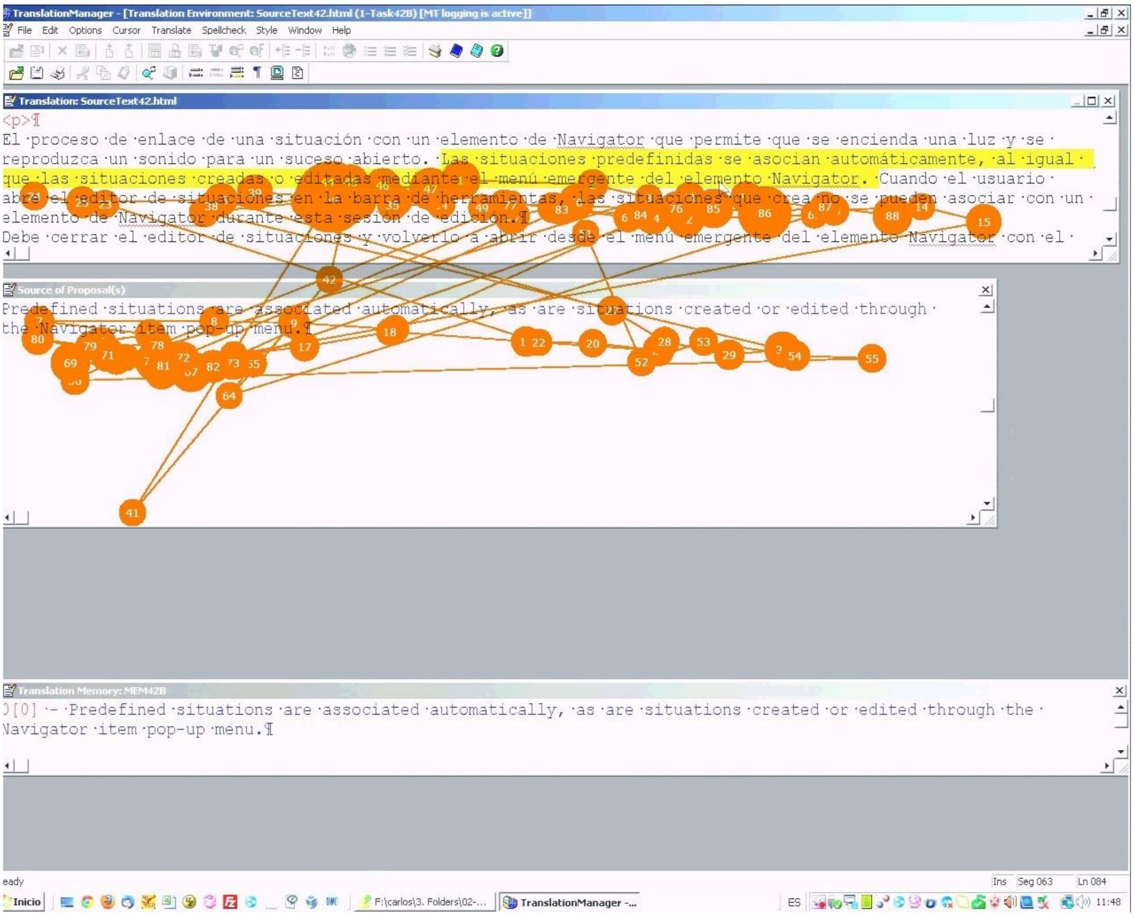
5.5.4. Example 4

In order to compare P02's behaviour in the Visual task, presented in Example 3, with his behaviour in the Blind task, let us take a similar segment, i.e. one with an almost perfect machine translation as its suggestion. The segment chosen for this comparison is segment 20 from SourceText42 and has the following characteristics:

> Source text in the active segment:
>
> Predefined situations are associated automatically, as are situations created or edited through the Navigator item pop-up menu.
>
> Translation suggestion:
>
> Las situaciones predefinidas se asocian automáticamente, **al igual** que las situaciones creadas o **editaras** mediante el menú emergente del elemento de Navigator.

The only difference between this suggestion and the translation that had been approved in the translated manual was “al igual”, instead of “igual” (both forms are equally valid), and “editaras”, which is a misspelling of “editadas”. Although “editaras” is not correct in this sentence, it exists as a verb form in Spanish, so it would not be detected as an error by a spell checker. The misspelling was not corrected by two translators working on the Visual task (P07 and P08) and by two translators working on the Blind task (P05 and P09). All translators kept “al igual”, except for one translator (P04) working on the Visual task, who misinterpreted the second clause and changed its meaning.

Figure 41. Gaze plot of P02 while translating a Machine Translation suggestion in the Blind task (Example 4)


TranslationManager - [Translation Environment: SourceText42.html (1-Task42B) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText42.html
Source of Proposal(s)
Translation Memory: MEM42B
eady
Ins Seg 063 Ln 084
Inicio
F:\carlos\3. Folders\02-...
TranslationManager -...
ES
11:48

<p>¶
El proceso de enlace de una situación con un elemento de Navigator que permite que se encienda una luz y se reproduzca un sonido para un suceso abierto. Las situaciones predefinidas se asocian automáticamente, al igual que las situaciones creadas o editadas mediante el menú emergente del elemento Navigator. Cuando el usuario abre el editor de situaciones en la barra de herramientas, las situaciones que crea no se pueden asociar con un elemento de Navigator durante esta sesión de edición.¶
Debe cerrar el editor de situaciones y volverlo a abrir desde el menú emergente del elemento Navigator con el

Predefined situations are associated automatically, as are situations created or edited through the Navigator item pop-up menu.¶

)[0] - Predefined situations are associated automatically, as are situations created or edited through the Navigator item pop-up menu.¶

Figure 41 presents the gaze plot for P02. There are several fixations around “editadas”, which was repaired correctly, and around “the Navigator item pop-up menu” (in the source) and “del elemento Navigator” in the target, where the translator deleted a preposition. The gaze plot shows a pattern similar to that presented in Figure 40, except that the translator did not look at the lower pane in search of metadata. The comparison suggests that when repairing suggestions from machine translation, the translator did not change his behaviour much because of the metadata.

5.5.5. Example 5

The last example to be examined here is taken from segment 15 in SourceText31. This segment also has a machine translation as its suggestion, but this time the suggestion has more problems to be fixed.

Source text in the active segment:

`"If you have not yet added an email address to your profile, click <span class=""ph menucascade""><span class=""ph uicontrol"">My IBM</span> &gt; <span class=""ph uicontrol"">Profile</span> &gt; <span class=""ph uicontrol"">Edit</span></span> and add it to your personal information."`

Translation suggestion:

Si aún no ha añadido una dirección de correo electrónico en su perfil, **haga clic en** <span class="ph menucascade"><span class="ph uicontrol">Mi IBM</span> &gt; <span class="ph uicontrol">**Profile**</span> &gt; <span class="ph uicontrol">Editar</span></span> y **añadirlo** a su información personal.

This segment is full of xml tags, indicated in grey above. The corresponding output would be:

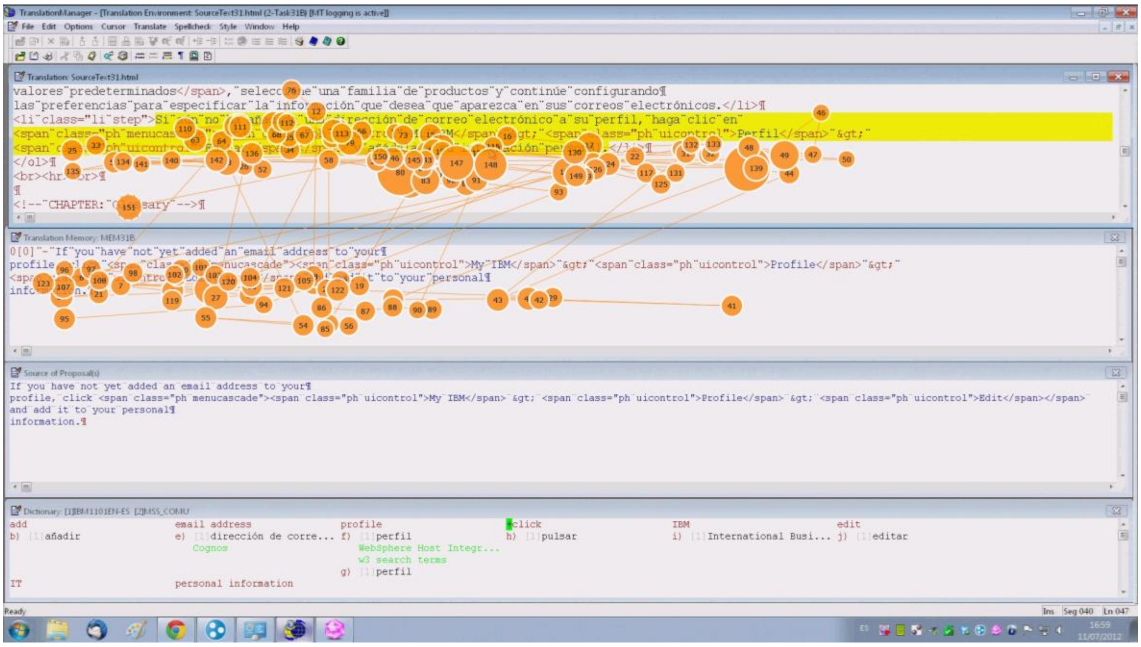
5. If you have not yet added an email address to your profile, click **My IBM** > **Profile** > **Edit** and add it to your personal information.

The parts indicated in bold refer to potentially problematic chunks in the suggested translation. The first one ("haga clic en") corresponds to the translation of "click". Although the translation is not incorrect, the reviewers were checking for internal consistency, so if the translator used one translation for "click" here, the same translation should have been used elsewhere in the file. Most translators changed the suggested translation to "pulse" (which was the recommended translation in the client's glossary, although consistency with a specific glossary was not being evaluated), except for P04 (working on Blind the task) and P06 (working on Visual the task). Those two translators failed to change "haga clic en" in this segment but translated "click" as "pulse" elsewhere, so they scored one error here. Therefore, the machine translation suggestion was responsible for internal consistency errors made by these two translators in this segment.

The second problem in the machine translation suggestion is the word "Profile", which remained untranslated. All translators fixed this problem correctly. This can be considered an easy fix, as even if a translator had missed it when working on the segment for the first time, the spell checker would have detected it as an unknown word. Finally, the third problem refers to the translation of "add it", which the translation suggestion presented as "añadirlo" (infinitive form with masculine pronoun) but should be translated as "añádala" (imperative form with feminine pronoun). All translators fixed the verb form correctly, but one translator (P09) failed to detect the wrong gender in the pronoun, which referred to "address" ("dirección", a feminine word in Spanish).

By watching the recording of P04, illustrated by means of the gaze plot presented in Figure 42, we see a work pattern that is similar to when this participant was translating a High Fuzzy Match in the Blind task. He alternates successively between the upper pane, where the pre-inserted suggestion is, and the middle pane, where the source text is available. He concentrates on translating "Profile" to "Perfil", then on correcting the verb form "añadirlo" to "añádala", and finally he replaces the preposition "en" with the preposition "a" in the first clause, making "Si aún no ha añadido una dirección de correo electrónico a su perfil,...". This was an optional change (the reviewers did not mark it as an error for those translators who failed to make the change), implemented by four translators (P01, P04 and P07, working on the Blind task; and P10, working on the Visual task). Although it was not a required change, it was a recommendable one, especially considering that the same verb "añadir" appears later in the sentence followed by the preposition "a" and no translator changed the preposition in the second occurrence to "en" to keep the linguistic consistency. Finally, P04 failed to change the translation for "click" from "haga clic en" to "pulse", which he had used elsewhere. (Note that this is also a term shown in the glossary in the lower pane, but the translator did not fixate there.)

Figure 42. Gaze plot of P04 while translating a Machine Translation suggestion in the Blind task (Example 5)


TranslationManager - [Translation Environment: SourceText31.html (2-Task31B) [MT logging is active]]
File Edit Options Cursor Translate Spellcheck Style Window Help
Translation: SourceText31.html
Translation Memory: MEM31B
Source of Proposal(s)
Dictionary: [1]IBM1101EN-ES [2]MSS_COMU
Ready
Ins Seg 040 Ln 047
16:59
11/07/2012

```
valores predeterminados</span>, selecc   e una familia de productos y continúe configurando¶
las preferencias para especificar la info   ción que desea que aparezca en sus correos electrónicos.</li>¶
```

```
0[0] - If you have not yet added an email address to your¶
profile,   <sp   clas     menucascade"><span class="ph uicontrol">My IBM</span> &gt; <span class="ph uicontrol">Profile</span> &gt;
```

```
If you have not yet added an email address to your¶
profile, click <span class="ph menucascade"><span class="ph uicontrol">My IBM</span> &gt; <span class="ph uicontrol">Profile</span> &gt; <span class="ph uicontrol">Edit</span></span>
and add it to your personal¶
information.¶
```

```
add                    email address               profile                   click            IBM                          edit
b) [1]añadir           e) [1]dirección de corre... f) [1]perfil              h) [1]pulsar     i) [1]International Busi... j) [1]editar
                          Cognos                      WebSphere Host Integr...
                                                      w3 search terms
                                                   g) [1]perfil
IT                     personal information
```

In complement to the points exemplified above, the available eye-tracking data allow the following general statements to be made:

- For Exact Matches, metadata helped participants translate faster, either because they looked at the metadata field and decided to not touch it (or to touch it minimally) or because the participants used a key combination to skip automatically the segments with an Exact Match as the translation suggestion (an *a priori* trust attribution).
- For (High and Low) Fuzzy Matches, metadata helped participants identify what parts of the sentence needed to be changed. This increased their translation speed, especially for High Fuzzy Matches, but at the same time added the risk of diverting the participants from finding other potential problems in the translation suggestions that were not indicated through the metadata.
- For Machine Translation suggestions, no distinguishable difference could be identified between the ways in which translators handled the translation suggestions under the two conditions. However, it can be assumed that the only piece of metadata available for this type of suggestion (the “m” indicator in the Visual task) might have affected the trust attributed to the suggestion, either positively or negatively, depending on the translators’ attitude towards machine translation.

Chapter 6. Discussion

Having presented the results in the previous chapter, I will now discuss the findings. I will start by testing the hypotheses set out in the Methodology chapter (section 4.2); I will then discuss other quantitative findings not directly related to the hypotheses, complement that discussion with the qualitative data from the interviews and finally summarise the findings.

6.1. Hypothesis testing

6.1.1. Hypothesis 1 (H1)

My first hypothesis states that: "The presence of *metadata* affects *translation time*". This hypothesis was confirmed by the statistically significant effect of Task (F = 32.71; p < 0.001), as presented in section 5.4.1.1. The statistical model estimates that the translators spent 54 seconds per 100 words more on the Blind task (without metadata) than on the Visual task (with metadata) on average, which corresponds to a difference of 43 percent between the two tasks. In other words, the absence of metadata was responsible for a 43-percent increase in Translation Time.

6.1.2. Sub-hypothesis 1a (H1a)

My first sub-hypothesis states that: "The effect of *metadata* on *translation time* varies in accordance with the *type of translation suggestion*". This sub-hypothesis was also confirmed, according to the results presented in section 5.4.1.4, where a significant interaction effect (F = 38.80; p < 0.001) was found between Task (metadata) and Suggestion Type.

Those results show that the absence of metadata was responsible for a 265-percent increase in Translation Time for Exact Matches (p < 0.001) and for a 24-percent increase in Translation Time for High Fuzzy Matches (p < 0.046). This is what one would expect, as metadata are supposed to help translators save time on identifying what needs to be changed in a given suggestion. The presence or absence of metadata did not significantly affect Translation Time for Low Fuzzy Matches and Machine Translations, probably because the changes required by these two suggestion types require a lot of time anyway.

6.1.3. Hypothesis 2 (H2)

My second hypothesis states that: "The presence of *metadata* affects *typing effort*". As explained in section 5.4.2, the results for Typing Effort were analysed in two steps, as the data distribution for this variable was strongly right-skewed, with a great concentration of data points at zero.

In the first step, Typing Effort was converted into a binary variable, so that the segments for which no edits were made received a value of "zero" and the segments that had at least one edit received a value of "one". After applying the statistical model assuming a binomial distribution on this dataset, a significant effect was found for Task (F = 6.266; p = 0.013), as presented in section 5.4.3.1, confirming the hypothesis. This result indicates that in the presence of metadata (the Visual task), translators edited 83 percent of segments (regardless of the actual amount of editing in each segment), whereas in the absence of metadata they edited 70.3 percent of segments. This is equivalent to saying that translators edited 18 percent more segments when metadata were present (Visual task) as compared to when metadata were absent (Blind task).

In the second step of the analysis, all the data points where Typing Effort was equal to zero were removed and the remaining data points were used in the statistical model, assuming a normal distribution after a logarithmic transformation. In this scenario, the statistical model found no significant main effect for Task (metadata).

These combined results mean that the presence of metadata is responsible for an 18-percent increase in the number of segments that are edited. This is not what one would expect, as metadata are supposed to help translators identify exactly what needs to be changed, but this result will make more sense when we look at the effect of metadata on the different suggestion types, in section 6.1.4 below. When only the segments with edits are considered, there is no significant difference in the number of edits between the two tasks (i.e. between the conditions with or without metadata).

6.1.4. Sub-hypothesis 2a (H2a)

The second sub-hypothesis states that: "The effect of *metadata* on *typing effort* varies in accordance with the *type of translation suggestion*". This sub-hypothesis was also confirmed, as in both steps of the statistical analysis a significant interaction effect was found between Task (metadata) and Suggestion Type. In the first step (F = 8.66; p < 0.001) (see Table 43), the statistical model shows that the presence of metadata is beneficial for Exact Matches, detrimental for High Fuzzy Matches, and on the verge of being significantly detrimental for Low Fuzzy Matches and Machine Translations. This reinforces the counter-intuitive result in the previous section, as translators produced more edits for both ranges of fuzzy matches and machine translation when metadata were available than when they were not available. An explanation could be that when metadata are not available, translators tend to overlook some necessary changes (and fail to implement them, thus reducing the amount of editing). One would expect that the failure to make the necessary changes would result in higher error scores, but this is not confirmed by the results (see sections 6.1.5 and 6.1.6 below). Further investigation is necessary to understand why this is the case.

In the second step of the analysis (see section 5.4.4), where only the segments with edits are considered, the interaction effect is also determined to be statistically significant (F = 3.224; p = 0.023), with some qualitative changes in the results due to the elimination of the segments with no edits.

6.1.5. Hypothesis 3 (H3)

My third hypothesis states that: “The presence of *metadata* affects *error scores.*” Similarly to what happened with Typing Effort, the data distribution for Error Score was strongly right-skewed, with a great concentration of data points at zero, which called for splitting the statistical analysis into two steps (see section 5.4.5).

In the first step, Error Score was converted into a binary variable, so that the segments with no errors received a value of “zero” and the segments with at least one error received a value of “one”. The statistical model was applied on this dataset assuming a binomial distribution and found no significant main effect for Task (F = 0.983; p = 0.322).

In the second step (see section 5.4.7), the segments with no errors were removed from the dataset, the remaining data points were log-transformed and the statistical model was configured assuming a normal distribution. Once again, no significant main effect was found for Task (F = 2.807; p = 0.095). This hypothesis is thus rejected.

These results are not intuitive, especially when metadata were found to affect the number of edits, so one keeps wondering how it is possible that a change in the number of edits does not imply a change in the number of errors. A possible explanation is that some of the edits that were made in the presence of metadata were not necessary changes (to avoid errors), but further investigation would need to be carried out to test this assumption.

6.1.6. Sub-hypothesis 3a (H3a)

The third sub-hypothesis states that: "The effect of *metadata* on *error score* varies in accordance with the *type of translation suggestion*." This sub-hypothesis is tested by looking at the result of the interaction effect between Task (metadata) and Suggestion Type. This effect was determined to be non-significant both in the first step (F = 0.349; p = 0.845) and in the second step (F = 1.148; p = 0.334) of the statistical analysis (see sections 5.4.6 and 5.4.7). Therefore, this sub-hypothesis is also rejected.

6.1.7. Conclusion

It is worth noting that, as explained in section 4.5.7.1, when metadata were present the translators had to insert the translation suggestion in the active segment, whereas when metadata were absent the suggestion had already been pre-inserted. In other words, it was not possible to isolate translation metadata as an independent variable, since it was always tied to the presentation mode (dynamic insertion vs. pre-insertion). This was compensated for in the statistical analysis by looking at Task as a variable (which combines translation metadata and presentation mode). With this reservation in mind, Table 62 summarises the results of the hypothesis testing.

Table 62. Results of the hypothesis testing

Hypothesis		Confirmed?	Result	Effect
H1	Metadata affects Translation Time	YES	When metadata were present, translators spent 125 seconds to translate 100 source words. When metadata were not present, translators spent 179 seconds to translate 100 source words. ⇒ Translators spent 43 percent more time on average when metadata were not present as compared to when metadata were present.	Shorter translation time
H1a	Suggestion Type interacts with Metadata	YES	The change in Translation Time due to the presence of Metadata depends on the Suggestion Type. When metadata were not present: ⇒ Translators spent 265 percent more time on average when dealing with Exact Matches. ⇒ Translators spent 24 percent more time on average when dealing with High Fuzzy Matches. There was no significant change in Translation Time for Low Fuzzy Matches and Machine Translation due to the presence of metadata.	E: Shorter translation time H: Shorter translation time L: No significant effect M: No significant effect
H2	Metadata affects Typing Effort	YES	When metadata were present, translators made changes in 83 percent of the segments. When metadata were not present, translators made changes in 70 percent of the segments. ⇒ Translators made changes in 18 percent more segments when metadata were present as compared to when metadata were not present.	Higher typing effort
H2a	Suggestion Type interacts with Metadata	YES	The change in Typing Effort due to the presence of Metadata depends on the Suggestion Type. When metadata were present: - For Exact Matches, translators edited fewer segments, but there was no significant difference in the amount of typing for those segments that were edited. - For High Fuzzy Matches, translators edited more segments, but there was no significant difference in the amount of typing for those segments that were edited. - For Low Fuzzy Matches, there was no significant difference in the number of segments edited, but the amount of typing was higher for those segments that were edited. - For Machine Translation, there was no significant difference in the number of segments edited or in the amount of typing for those segments that were edited.	E: Lower typing effort H: Higher typing effort L: Higher typing effort M: No significant effect
H3	Metadata affects Error Score	NO	There was no significant change in Error Score due to the presence or absence of metadata.	No significant effect
H3a	Suggestion Type interacts with Metadata	NO	There was no significant change in Error Score due to the interaction between metadata and Suggestion Type.	No significant effect

6.2. Additional quantitative findings

In addition to providing the basis for testing my three sets of hypotheses, the statistical analysis presented in section 5.4 was of an exploratory nature. It included not only the main independent variables that are contained in the hypotheses – metadata (Task) and type of translation suggestion (Suggestion Type) – but also several additional independent variables (see Table 17). This exploratory analysis resulted in some interesting findings, particularly for Translation Time, which by its very nature could be analysed in a single step and produced a high number of significant interactions. The following sections will cover the findings based on the independent (explanatory) variables, some of which have also been presented in section 6.1 above.

6.2.1. Metadata

Translation metadata were represented by the variable Task in the statistical analyses, where the Visual task corresponded to the presence of metadata and the Blind task corresponded to the absence of metadata. Metadata were found to produce significant effects on translation time and typing effort in opposite ways: in the presence of metadata, the translators spent significantly less time translating the documents, but they invested a significantly higher typing effort. Among other things, this result suggests that typing is not what determines the amount of time spent during translation. One can type less when performing a specific task and still spend more time on it, which indicates that time is being invested in activities other than typing (e.g. deciding between different translation solutions).

Metadata were not found to produce a significant effect on Error Score. In other words, the information about the translation suggestions did not help translators make fewer errors, nor was it responsible for an increase in the number of errors.

These combined findings suggest that translators worked in the following ways:

- When they had metadata available, translators spent little time identifying what they had to change (if anything) and implemented all or most of the necessary changes. They used different strategies according to the type of translation suggestion (no changes for exact matches, many changes for low fuzzy matches and machine translation).
- When they had no metadata available, translators took longer to identify the type of suggestion they were being offered and to decide on the type of editing

strategy they needed to apply to repair the translation suggestion. However, they seem to have overlooked some of the required changes, which is what might explain the fact that they edited the suggestions less frequently than in the presence of metadata.

- Due to the good quality of the translation memory and of the machine translation engine, any failures to edit the translation suggestions when required did not impact on the error scores significantly.

It remains to be explained why in the presence of metadata (Visual task) the translators edited 83 percent of the segments, when we know that 25 percent of segments had an Exact Match as their translation suggestion (which in principle requires no edit). I will return to this later (see section 6.2.3).

6.2.2. Suggestion Type

Suggestion Type is the only independent variable that proved to produce significant effects on all three dependent variables. Exact Matches had the lowest translation times, the lowest typing effort and the lowest error scores of all suggestion types. The three other suggestion types ranked differently for each of the dependent variables. Table 63 summarises the ranking of each of the four levels of Suggestion Type on the three dependent variables.

Table 63. Effect of Suggestion Type on the three dependent variables

Dependent variable	Ranking of the different levels of Suggestion Type
Translation Time (seconds / 100 words)	E < H < L = M
Typing Effort (%)	E < H = M < L

Error Score (errors / 100 words)	E = L < H = M (binary)
	E = H < L = M (non-zero)

Legend:
E: Exact Matches
H: High Fuzzy Matches (85-99%)
L: Low Fuzzy Matches (70-84%)
M: Machine Translation

Looking at Translation Time first, Exact Matches required the lowest translation times. It is plausible to assume that this happened because Exact Matches were quickly identified as good suggestions by the translators – not only in the Visual task, where they were indicated as exact matches, but also in the Blind task. The same can be said about High Fuzzy Matches, which were the second fastest suggestion type to be translated.

When it comes to Low Fuzzy Matches and Machine Translation, these suggestion types required the longest translation times. This is an expected result for Low Fuzzy Matches, as they are the suggestion type that normally requires the most changes among the translation-memory matches. Machine Translation can require a higher or lower number of changes depending on the quality of the engine, but it can also activate strategies that are different from (and possibly more time-consuming than) those used for dealing with translation-memory matches.

Next, if we look at the results for Typing Effort in Table 63, Exact Matches required the lowest typing effort, Machine Translation and High Fuzzy Matches required an intermediate level of typing effort, and Low Fuzzy Matches required the highest typing effort. Looking at these results from the perspective of Machine Translation, we see that it required a level of typing effort similar to that of High Fuzzy Matches but a lower typing effort than that of Low Fuzzy Matches, even if it required the most time to process, as indicated in the previous paragraph. A possible interpretation for this is that even though translators needed more time to process Machine Translation suggestions they still failed to make some of the required changes (reflected in the intermediate typing effort), thus producing more errors, as Machine Translation had the highest error scores, as will be seen below.

For Error Score, Suggestion Type produced different results depending on whether Error Score was analysed as a binary variable (how many segments contained errors, regardless of the number of errors) or as a regular variable, taking into account only those segments that contained errors (how many errors they had). To clarify things, the errors “contained” in a segment refer to the errors identified in the final translated segment after the translator worked on it, as detected by the reviewers; this has nothing to do with potential errors in the source text or in the initial translation suggestion.

Exact Matches ranked best in the two types of analysis, while Machine Translation ranked worst. For High Fuzzy Matches and Low Fuzzy Matches, the performance of these two types of translation suggestions changed depending on how the variable was analysed: not many segments that contained a Low Fuzzy Match as the translation suggestion produced errors (together with Exact Matches, this was the suggestion type with the lowest ratio of erroneous segments), but when those segments did produce errors, they produced many errors (together with Machine Translation, this was the suggestion type with the highest error scores among the erroneous segments). For High Fuzzy Matches, the opposite happened: many segments that contained a High Fuzzy Match as the translation suggestion produced errors (together with Machine Translation, this was the suggestion type with the highest ratio of erroneous segments), but when those segments did produce errors, they produced relatively few errors (together with Exact Matches, this was the suggestion type with the lowest error scores among the erroneous segments). This can be explained by the fact that the required changes in High Fuzzy Matches (due to slight differences between the two source texts) are more difficult to spot than the required changes in Low Fuzzy Matches, especially in the Blind task; however, even when the required changes are not spotted, the errors produced are not very numerous. Conversely, the required changes in Low Fuzzy Matches are more obvious, but when they are not implemented, the errors produced are more numerous.

In the current section, we have looked at the effect of Suggestion Type regardless of the presence or absence of translation metadata. The translators' strategies that might have led to these results will be interpreted again in the next section, when analysing the interaction between Suggestion Type and Task.

6.2.3. Metadata and Suggestion Type (interaction)

The interaction effect between Metadata (Task) and Suggestion Type proved to be significant for Translation Time and Typing Effort, but not for Error Score.

For Translation Time, metadata were beneficial for Exact Matches and High Fuzzy Matches, since the absence of metadata resulted in a sharp increase in translation times for those types of matches (of 265% and 24%, respectively). For Low Fuzzy Matches and Machine Translation, there was no significant difference in translation times between the two conditions (see Table 27). When metadata were available, Low Fuzzy Matches and Machine Translation required virtually the same translation time, High Fuzzy Matches required approximately 33 percent less time than both suggestion types, and Exact Matches required 74 percent less time than High Fuzzy Matches. When metadata were not available, High Fuzzy Matches, Low Fuzzy Matches and Machine Translation required virtually the same time and Exact Matches required approximately 30 percent less time than those three other suggestion types (see Table 28).

In the case of Typing Effort, the variable had to be analysed in two steps due to its data distribution (see section 5.4.2). The first step looked at the percentage of segments that were edited (at any rate). This analysis showed that translation metadata were beneficial for Exact Matches (there was a 153-percent increase in the number of edited segments when no metadata were available) and detrimental for High Fuzzy Matches (there was a 26-percent decrease when no metadata were available). For Low Fuzzy Matches and Machine Translation, the availability of metadata actually seems to have increased the number of edited segments, but the statistical results are on the verge of significance, with $p \approx 0.05$ (see Table 43).

When metadata were available, the rate of edited segments when the suggestion type was a High Fuzzy Match, a Low Fuzzy Match or a Machine Translation was between 87 and 97 percent, with no statistical difference between those suggestion types, while only 15.6 percent of Exact Matches were edited (see Table 42). In principle, we could expect authentic exact matches to require no edits, fuzzy matches of any level to require at least some editing, and machine translation to require edits depending on the quality of the engine setup. The 15.6-percent rate of edited Exact Matches (as opposed to the expected zero percent) indicates that (some) translators did not trust the translation memory entirely or that the translation memory actually contained errors that needed to be fixed. The lower-than-100-percent edit rate for High Fuzzy Matches and Low Fuzzy Matches indicates that, in some segments, translators overlooked the metadata information or considered that the changes required in the segment according to the metadata were actually unnecessary or irrelevant.

When metadata were not available, Exact Matches were still the suggestion type with the lowest edit rate, but the difference between Exact Matches and the other suggestion types was reduced. Because translators did not know the type of suggestion they were editing, they edited Exact Matches more often than they would have if they had had this information; they edited the other two types of translation-memory suggestions less often than when they knew what had to be changed; and they edited machine translation less often than when they knew they were editing machine translation. The results for low fuzzy matches and machine translation have a p-value of 0.05, so the assumptions on these two types of suggestion cannot be made with full certainty. If these results were confirmed with further testing, they would reflect a mistrust of machine translation, indicating that translators tend to accept suggestions more often when they do not know they come from machine translation.

The second step in the analysis of Typing Effort looked at the variable in its original meaning (the percentage of characters typed in relation to the total number of characters in the final target segment), but taking into consideration only those segments where at least one character was typed. In this case, metadata only affected Low Fuzzy Matches, which required a much lower typing effort when metadata were not available (see Table

49). When metadata were available, Low Fuzzy Matches required the highest typing effort (37.2%), followed by High Fuzzy Matches and Machine Translation, and followed by Exact Matches, which required only 3.3% of typing effort. When metadata were not available, the differences between the four suggestion types were reduced: Exact Matches and High Fuzzy Matches required the lowest typing effort (~ 7%), while Low Fuzzy Matches and Machine Translations required the highest typing effort (~ 15%) (see Table 48 and Table 50).

From the perspective of translation metadata, my findings can be summarised as follows. When translators did *not* have access to metadata, the following things happened:

- For Exact Matches, they spent 265% more time, edited 153% more segments and made the same number of errors;
- For High Fuzzy Matches, they spent 24% more time but edited 26% fewer segments and made the same number of errors;
- For Low Fuzzy Matches, they spent the same time, edited the same number of segments (although the amount of typing was much lower) and made the same number of errors;
- For Machine Translation, they spent the same time, edited the same number of segments and made the same number of errors.

It is surprising that translation metadata did not affect translation quality for any of the suggestion types. That is, the suggestion types have their inherent error scores (see the previous section), with Exact Matches ranking best and Machine Translation ranking among the worst, but these were not affected by the presence or absence of metadata. One could expect that seeing what has to be changed in a translation suggestion (based on the translation metadata provided in the Visual task) would help to reduce the number of errors in the final segment, but my results show that any differences in quality between the two conditions are not statistically significant, at least for the quality requirements that are common in the type of translation project used in the experiment.

6.2.4. Text

The Text variable was found to produce a significant effect on Translation Time: the translators took 24 seconds more on average to translate 100 source words of Text42 than to translate 100 source words of Text31, a difference of 17 percent (see section 5.4.1.3). Both source texts were taken from the same software manual and I took pains to make them as similar as possible, even at the segment level (see section 4.5.5). However, as indicated by the results for this effect, the two texts cannot be considered fully equivalent. It would be interesting to carry out further investigation by looking into individual segments in relation with individual participants to try to speculate on potential reasons for this surprising result.

In any case, in anticipation of a potential difference between the source texts, both texts were assigned to each task type (Blind or Visual) and to each task order (first task or second task) alternately, as indicated in Table 8. That is, Text31 was translated in the Blind task by five translators and in the Visual task by five translators, and the same happened with Text42. Likewise, Text31 was translated first by five translators and second by five translators, and the same thing happened again with Text42. This strategy of alternating the texts was meant to neutralise any potential effects of the Text variable. Moreover, the analysis for the remaining variables already takes into account these potential variations, since the statistical model is also adjusted by text.

Despite the difference between the texts in terms of the time required for translation, it is worth noting that the two texts required similar typing efforts and produced similar error scores, as the statistical model found no significant main effect for Text on these variables.

6.2.5. Gender (not significant)

The participants in my experiment were evenly distributed as far as gender is concerned: five men and five women. According to the findings presented in the Results chapter, no significant differences were found between men and women for any of the dependent variables (no significant main effect for Gender according to the linear mixed effects model). However, the interaction effects between Task and Gender ($F = 18.69$; $p < 0.001$) and Suggestion Type and Gender ($F = 11.53$; $p < 0.001$) on Translation Time were determined to be significant (see sections 5.4.1.5 and 5.4.1.6). Two phenomena might be responsible for these results.

The first phenomenon is that three female participants (P03, P05 and P09) spent very little time translating Exact Matches, while this is not the case for women in the Blind task nor for men in any of the tasks. Going back to the keystroke logging data and the video recordings, I noticed that those three participants used a key combination (shortcut) in the translation tool that automatically skipped exact matches, so that they did not even activate a segment for translation when its suggestion was an exact match. This strategy would only work when the tool was configured for the Visual task, since in the Blind task there were no suggestions coming from the memory (they had already been pre-inserted in the target segments). In other words, it was not about seeing the translation suggestion and deciding to accept it immediately because it was an exact match; that decision had implicitly been made beforehand when choosing to use this key combination systematically. It would be necessary to carry out further studies to try to find out why those three participants decided to ignore exact matches in the first place while the others did not, and whether gender played a non-random role in this decision.

The second phenomenon is that one male participant (P08) was particularly slow in the translation tasks, especially in the Visual task, despite his long experience as a translator. In the interviews, he said he had become nervous because of the experimental setting ("white-coat effect"). As we shall see below, his unusual performance, combined with the extreme opposite strategy used by some women mentioned in the previous paragraph, was responsible for some of the significant interactions of Gender as an explanatory variable.

6.2.6. Metadata and Gender (interaction)

As just mentioned above, Gender in itself was not determined to be a significant main effect. The non-significant difference between men and women remains if we look at each of the tasks individually: even in the Visual task, where there was a great difference (77 percent) in the means for Translation Time between men and women (see Table 31), the difference was not determined to be statistically significant.

Yet Task (metadata) was found to be a significant main effect on Translation Time (see sections 5.4.1.1 and 6.1.1). The interaction between Task and Gender shows that when the women's performance is compared between the two tasks (see Table 30), women spent significantly more time (79 percent) on the Blind task (no metadata) as compared to the Visual task, while there was no significant difference between the tasks for men.

These dissimilar results between men and women can be explained by the performance of participant P08, who spent much more time on the Visual task than on the Blind task, as explained in the previous section. If the data for this participant are removed from the analysis, the difference between the Visual and the Blind tasks for men also becomes significant, with $p = 0.021$, and the main effect of Gender remains non-significant.

To sum up the results of the interaction between Metadata and Gender:

- Women had significantly lower translation times in the presence of metadata than in the absence of metadata.
- Men also had lower translation times in the presence of metadata, but the difference between the two tasks is only significant if we remove P08, who was extremely slow in both tasks, and especially in the Visual task.
- Women had lower translation times than men did in the presence of metadata (although the difference is not significant, with p = 0.088), especially due to the way some women handled Exact Matches.

6.2.7. Suggestion Type and Gender (interaction)

The interaction between Suggestion Type and Gender on Translation Time was also determined to be significant. The strategy used by three female participants of jumping the exact matches through an automatic key combination, as explained in section 6.2.5, is the main reason for this difference. As indicated in Table 34 in section 5.4.1.6, there is a significant difference in Translation Time between men and women when translating exact matches. Although the table also suggests that translation times are consistently lower for women than for men across all suggestion types, the differences for the other suggestion types were not determined to be significant. Moreover, these mean differences are reduced when the male participant P08 is removed from the analysis.

As presented in the Results chapter, Exact Matches required the lowest translation times for both genders, while Low Fuzzy Matches and Machine Translation required the highest times (non-statistically significant difference between the two). The only gender-related difference is that women spent less time translating High Fuzzy Matches than Low Fuzzy Matches or Machine Translation, whereas for men there was no significant difference between High Fuzzy Matches, Low Fuzzy Matches and Machine Translation.

To sum up the results of the interaction between Suggestion Type and Gender:

- Women spent significantly less time than men when translating Exact Matches, especially because three women skipped this suggestion type automatically using a keyboard shortcut.
- Women also spent less time than men when translating the three other suggestion types, but the mean differences in this case were not statistically significant.

6.2.8. Text and Experience (interaction effect)

The last effect that was found to be statistically significant was the interaction between Text and Experience on Translation Time. To recall, high experience corresponds to translators with more than five years of work experience, while low experience corresponds to translators with less than five years of work experience, and there are five translators in each category. Text had already been determined to be a significant effect in itself, as Text42 was found to require 17 percent more time to translate than Text31 (see sections 5.4.1.3 and 6.2.4). Now the interaction between Text and Experience shows that the more experienced translators spent significantly more time on Text42 than on Text31 (while the time difference between the two texts was not statistically significant for less experienced translators). Further investigation is necessary to understand this phenomenon.

6.2.9. Non-significant effects

The relevance of the findings from the statistical tests lies not only in the significant effects that were found but also in the non-significant ones. For example, one might expect Experience to play a major role in translator performance, but my results show no significant effect of Experience on any of the dependent variables (except in an interaction with Text, as just explained). It is worth recalling, however, that the least experienced participant in my experiment had 1.5 years of full-time work experience as a translator, so my study differs from those that compare novices (usually final-year students or translators with less than one year of work experience) with professionals. In my case, the comparison was between five translators with 1.5 to 3.5 years of experience and five translators with 7 to 18 years of experience, and no significant differences were found between the two groups.

Another factor that was included in the analysis is Age. This factor had a strong correlation with Experience, and since Experience did not produce significant effects, there was no reason to expect that Age should produce significant effects either. If Age were to influence the results despite the non-significant results for Experience, then we would need to look into other factors related to age – besides the building up of professional experience – that could have an influence on performance. These might be, for example, language views (older translators could be more conservative with respect to language structures or less tolerant of certain types of errors) or computer literacy (one could expect older translators to be less – or more – proficient in the use of technology). However, the non-significant results for Age suggest no such phenomena.

Task Order was also taken into account in the analysis, as I feared that the order in which the translators had to perform the Visual or the Blind tasks could affect their results. For example, one might expect that participants would be more tired in the second task, which would increase translation times or error scores, or, inversely, one might hypothesise that in the second task translators would be more relaxed and more familiar with the experimental setting, and would thus perform better. Neither of these hypothetical assumptions proved to be true, and the order in which the tasks were presented did not affect the results significantly.

Similarly, Text Order was also included in the analysis, to test for potential changes in the results due to the different order in which the texts were presented. Once again, the factor was not found to produce any statistically significant effects. We should bear in mind that all participants were used to translating, revising or post-editing thousands of words a day, and the word volume in all my four tasks combined (the two preliminary tasks and the two main tasks), of around 1,200 words, lay within the range of their regular work load.

Finally, three other independent variables were included in the analysis and produced no significant results: Copy Time, Copy Effort and Copy Errors. These variables correspond to the translators' performance in the preliminary Copy task, which consisted of typing on the computer a text that was presented on paper, and were used as baseline measurements to check whether, for example, translators that copied faster also translated faster. The underlying idea is that the Copy task does not include the translational component in its cognitive processing, i.e. it isolates the typing skills, measuring how much time the translators spend producing a certain number of words, how many corrections they make and how many errors they overlook, when they do not have to think about producing content (original writing) or about interlingual or intercultural differences (translating). Interestingly, the data showed no significant correlation between these indicators and the participants' performance when translating. Further, there was virtually no correlation between translating from scratch and translating with the help of translation suggestions (except for a marginally significant correlation for error scores between the Scratch and the Visual tasks) (see section 5.1.4).

It is plausible to make two assumptions based on these findings. First, that other cognitive activities in the translation process take much more time than typing does, and since typing represents a small percentage of the total time spent on translating, it is not an important factor in the overall translation time. (Similarly, the typing effort invested when translating is due not only to a linear flow of text production, but also to deletions, corrections and word replacements that are typical of the translation activity – and the same applies to the types of errors that are made.) Second, that the translation tools provide relevant information (suggestions and metadata) that help translators find and choose (respectively) different translation solutions. Therefore, the more efficient translator is the one who knows how better to use the information provided by the tools, thus making better decisions, and this overrides any differences based on typing skills. Similar findings are reported in other studies (cf. Krings 2001; Sharmin et al. 2008). It is also interesting to see what translators think about the topic. Some think touch-typing is a really important skill to learn, while others ponder that typing speed is not at all important. Those different opinions can be found in translators' forums, such as in the discussions opened by Moran (2014) and McKay (2011).

6.3. Interview data

The interview data provided the qualitative component of my research. They allowed me to obtain a more human view of the translation process, from the point of view of the participants themselves. This material does not provide information on tasks at the segment level, since one cannot expect human memory to retain the information on every segment in the three translation tasks (Scratch, Visual and Blind). Therefore, the verbal data from the interviews can only be *compared* with the results about the whole texts in each task, as presented at the beginning of Chapter 5 (sections 5.1 to 5.3), making no distinction about the suggestion types within the texts. As a reminder, the Scratch task contained no translation suggestions, while the Visual and Blind tasks offered translation suggestions distributed in similar ways, with seven suggestions of each of the four types. The difference between the Visual and Blind tasks lay only in the presence of metadata in the Visual task and in the pre-insertion of suggestions in the Blind task.

As far as *metadata* are concerned, the quantitative results presented in section 5.1 do not show any advantage in terms of performance in favour of a specific translation task when the texts are considered as a whole. However, the testimonials gathered in the interviews and presented in section 5.2 indicate that the participants tended to believe they performed better in the more traditional Visual task – with translation suggestions and metadata –, as there was a general propensity to over-rate that task. The feeling of enhanced performance might help explain why most participants preferred to work on the Visual task, but another factor that played a prominent role for this preference was *task familiarity* (and the increased level of confidence resulting therefrom). In sum, the translators considered the task they were more familiar with to be faster, more comfortable and able to give better quality than the other tasks.

The information collected in the interviews indicates that metadata were also a relevant factor for increasing confidence and comfort, by giving translators a hint on how to initially approach a suggestion. That is, the translators reported using different strategies for different types of suggestions, and metadata helped them identify the suggestion types they were working with.

Another factor that proved to affect translators psychologically in the way they approached the text and the trust they attributed to the suggestions is *pre-translation*, i.e. the pre-insertion of translation suggestions in the Blind task. The participants often talked about the “translation” (referring to the Visual task) as opposed to the “revision” (referring to the Blind task), where they perceived the text as having being previously translated by an (assumedly reliable) human translator. Since both tasks had the same kinds of translation suggestions, what seems to have caused this misperception is that the participants associated the Blind task, where the segments came pre-translated, with the revision tasks they were accustomed to doing at the company, where the text would have already gone through a translation phase similar to the Visual task. An unintended difference between the instructions for both tasks could also have contributed to this distinction, as mentioned in section 4.5.7.3.

6.4. Summary

Recapitulating the results of testing my six hypotheses and sub-hypotheses, four of them (H1, H1a, H2 and H2a) were confirmed and the other two (H3 and H3a) could not be confirmed (see Table 62). Testing of H1 shows that metadata affected translation times, i.e. the translators had lower translation times overall when metadata were available (in the Visual task). Testing of H1a shows that this effect of metadata on translation times varies according to the types of translation suggestions, i.e. Exact Matches had a strong reduction in translation times when metadata were available, High Fuzzy Matches had a smaller reduction, while Low Fuzzy Matches and Machine Translation were not affected by the presence of metadata. Testing of H2 shows that metadata also affected typing effort, but in this case, the presence of metadata produced an overall increase in typing effort. According to H2a, the effect of metadata on typing effort also depends on the type of translation suggestion, as the presence of metadata reduced the typing effort for Exact Matches, increased it for High Fuzzy Matches and Low Fuzzy Matches and had no significant effect for Machine Translation. Finally, the presence of metadata was not found to affect error scores (H3), not even when the different types of translation suggestions are considered individually (H3a).

Thinking in terms of productivity, projects involving the leverage of translation memory matches with a high percentage of exact matches and high fuzzy matches should favour translation environments that present metadata, as metadata have proved to reduce translation times for those types of matches, with no impact on translation quality. A doubt remains as to whether the gain in productivity detected for high fuzzy matches might happen at the expense of a greater cognitive load, because of the increased typing effort. However, as discussed elsewhere (see sections 5.2.1.2 and 7.4.3), there is no clear indication that typing effort correlates with general cognitive load, especially considering that professional translators are used to typing a lot.

For projects that contain a high percentage of segments with machine translation and low fuzzy matches, my results give no reason to support the use of an environment with metadata, as this factor did not improve speed or quality, and caused an increase in typing effort when dealing with low fuzzy matches. However, the metadata for machine translation was virtually inexistent in both tasks (except for an indication that some translation suggestions came from machine translation, in the Visual task) and we do not know the impact that additional metadata elements for machine translation might have on translators' performance.

In sum, translation metadata reduced translation times with no significant impact on error scores, although they increased typing effort for some suggestion types. Considering that in the interviews the participants preferred the task with metadata and did not perceive the increased typing as uncomfortable, it seems plausible to recommend the use of metadata as a general strategy for workflows that combine translation memories and machine translation.

Chapter 7. Conclusion

7.1. Findings

The main research question in this thesis is whether and how translation metadata affect translators' performances. This question has been answered by showing that translation metadata do affect translation time and typing effort, and that the effects vary according to the type of translation suggestion (exact matches, fuzzy matches, machine translation). The qualitative data obtained in the interviews have shown that translators also mentioned metadata as a helpful feature in the translation tool, among other reasons because metadata help them adapt their translation strategies more easily according to the suggestion type.

The interviews have also shown that translators' perceptions are affected by the way the texts are presented (either in "dynamic" mode or in "pre-translation" mode). Translators tend to trust more those translation suggestions that have been produced by a peer (a human translator, but also through a translation memory), and they tend to mistrust what comes from machine translation, even when they recognise that the engine used in the company tends to present good suggestions overall. This might be explained by the fact that the translators are much less familiar with machine translation than with translation memories and traditional revision processes, and by their lack of knowledge of the technology behind statistical machine translation (as the engines are usually trained on the basis of translation memories). This can also be attributed to a generalised mistrust of machines as compared to humans. It is worth noting that the increased familiarity with translation memories over the years has allowed them to be perceived as much more "human" than was the case in the past, so the same phenomenon might be expected to happen with machine translation in the future.

As a complementary finding, the current study identified no significant correlation between the translators' performances while typing and their performances while translating. This result reflects the fact that the translation process involves many more cognitive activities than just typing, as reflected in performance indicators such as time, edits and errors. For example, in terms of time, since typing represents a small percentage of the total time spent on translating, it becomes a less important factor in the overall translation time. Therefore, the types of aids that seem to help translators the most in achieving higher overall performances while translating are those that help them save "thinking time" rather than "typing time". Moreover, those results also suggest that the translation tools used in the translators' daily work tend to homogenise their performance.

In sum, translation suggestions and the associated metadata are most useful not for what they save translators in terms of typing, but for the way they help them find, choose and implement translation solutions.

7.2. Applicability

I hope this study has contributed to the knowledge of translation and post-editing processes and can help to improve workflows and practices. This increased knowledge can benefit all parties involved in the translation scene, from translators to translation companies, translation-tool manufacturers, translation customers and translation users.

One of my concerns has been to explore possibilities of how to optimise the translation process in ways that could help increase not only productivity (and earnings) but also job satisfaction among translation professionals. On the other end, besides the potential impact on costs, the search for optimal processes can increase the volume of text that can be processed, which is a gain for translation buyers and users.

If some recommendations can be made for best practices in the industry, my results indicate that the best workflow would be to provide translators with an environment with metadata, leveraging translation memory matches above 85% (even though slightly lower percentages should still be tried) and replacing the remaining segments with machine translation, provided that the engine is of appropriate quality and provides acceptable terminology consistency. If metadata for machine translation are available, they have the potential to produce positive effects as well, although this possibility has not been tested.

7.3. Contribution to the field

Besides the more practical applications of the body of knowledge produced in this thesis, I hope the results are also of intellectual importance, as I have found that the impact of technology lies not just in what it does, but also in what the stakeholders *think* about what it does.

I have found that items such as metadata and presentation mode (in my case, "dynamic" vs. pre-translated) are relevant factors to be taken into account when analysing the results of studies on the translation process. This finding should warn us against making simple comparisons between studies that are different in this respect. For example, reports such as those presented in Autodesk (2011) seem to compare machine translation with fuzzy matches in a post-editing environment (with no metadata). This comparison could introduce a bias in favour of machine translation, as fuzzy matches are not normally used without metadata, and there is no need to use them in this way. I have shown that metadata affect performance indicators in several ways, and should therefore advise that researchers be aware of this when designing future experiments.

I hope this thesis has also contributed to the field in terms of the methodology of workplace studies, by presenting some challenges and solutions. An important lesson is the need to find an optimal balance between ecological validity and data validity when conducting translation experiments in realistic scenarios.

7.4. Limitations of the present study

Like most studies that try to obtain a detailed view of the translation process, the current study involved a limited number of participants (ten) with a certain level of expertise (professionals), one language combination (English into Spanish), one text type (software manual), one translation tool (IBM TranslationManager), and so on. Even though one might be tempted to generalise the findings of individual studies to the translation process in general, it is only the combination of several studies – even if having similar limitations – that can allow us to think that some conclusions are of general nature.

In the following sections, I will list a number of shortcomings that were specific of my experiment, in addition to those general limitations related to translation process research.

7.4.1. The observer's paradox

O'Brien (2009: 253) points out a basic challenge in translation process research: "we wish to observe what professional translators 'normally' do, but we remove them from their 'normal' work environments in order to do so", referring to the "Observer's Paradox" (Labov 1972: 209). In order to compensate for such contradictions, reproducing a real-world environment was a top priority in my study. The increase in ecological validity came together with an increase in the complexity of the experimental set-up. This involved not only choosing appropriate source texts, preparing authentic translation memories and allowing translators to work with a translation tool they were familiar with, but it also meant taking the research tools to the translators' computers, instead of having translators come and use a computer set up specifically for the experiment in a more controlled setting. Thanks to this approach, the translators were allowed to keep using the computer they were most accustomed to, including their hardware (screen monitor, mouse, keyboard) and software (operating system, glossaries, browsers, quality assurance tools), together with all configurations (shortcuts, colours, etc.).

My involvement at the workplace over a period of four months brought several benefits, such as the possibility to observe some aspects of their behaviours not directly related to the experiment. Most notably, through the contacts I established with the company's translators in informal situations (e.g. during coffee breaks and lunches), and by the time the experiments took place – two months into the secondment period – it is safe to assume that the participants felt more relaxed in my presence and trusted me more than if they were seeing me for the first time. Similar benefits of the continued presence of the researchers at the workplace are also reported by Ehrensberger-Dow (2014: 368). Notwithstanding all the increased confidence between the participants and the researcher, it is also plausible to assume that they were not working as naturally as they would have been if they were not being observed.

The "threat" posed by the presence of the researcher – either physically at the moment of the experiment (at a nearby desk) or later on when analysing the results – is a known factor that disturbs the naturalness of the workplace, known as the "white coat effect" (O'Brien 2009a: 258–259). Another potentially intimidating factor is the research equipment, especially the eye tracker. Although some participants mentioned this issue in the interviews, most said that they were thinking about the eye tracker only at the beginning of the experiment (the Copy task) but then forgot about it. P08 was the only participant who reported having become somewhat nervous because of the research equipment during the whole experiment:

> Researcher: But during the experiment, were you thinking very often that you had... [the eye tracker in front of you]?
>
> P08: Now that I think of it, yes, yes...
>
> Researcher: During the whole time?
>
> P08: Yes. At least I'm conscious of it... Some people are capable of... can ignore it... I can't.

This might help explain why P08 was noticeably slower than the other participants and made so many changes to his translations, even though he had more than 20 years of experience working with the same materials and tools as those used in the experiment.

7.4.2. Quantitative data collection methods

While the use of keystroke logging and video recording tools worked as expected and did not pose major problems, eye tracking proved to be a more challenging data collection method. Issues with inaccurate and inconsistent calibration and with incomplete capturing of the screen image were explained in section 4.6.3 and illustrated in section 5.5. Apart from those, other aspects of the ecological validity I was trying to achieve posed difficulties. One aspect is that the translators worked too close to the screen monitor (and consequently to the eye tracker) as compared to the distance of approximately 70 cm recommended by the eye-tracker manufacturer (Tobii Technology 2008: 11). When this happened, the translators were asked to move the monitor a little further away, until reaching a minimum distance of 60 cm, which meant they had to work at a distance greater than their normal working distance. After repositioning the monitor and the eye tracker, there were cases where hands, arms or eyewear frames sometimes stayed between the participant and the eye tracker, obstructing the infrared beams and compromising the eye-tracking data. According to my estimates, data losses of up to 30 percent should be expected in such studies – because translators naturally look away from the screen at times (e.g. to look at the keyboard or mouse, or while thinking). However, for one participant (P03) I had a data loss rate of up to 65 percent, and her data had to be used with extreme care (see Figure 34 in section 5.5.1).

Since eye tracking was not of utmost importance in my research design, the solution to those problems was to use the eye-tracking data available only where they could actually provide reliable information. Therefore, the data were used for two main purposes:

- to increase the precision of the time measurements per segment, by detecting the area of the screen the participant was looking at when a confirmation was needed on whether they had already started to work on a segment (or were still working on a segment);
- to identify the participants' strategies during the translation of specific segments or structures, as illustrated in section 5.5.

If a study uses eye tracking as a major data collection method, then other compromises must be found, probably by selecting participants in a different way or by giving up on some of the ecological validity in favour of a higher precision in the data, such as setting up a single computer for all the participants with a monitor-mounted eye tracker. Even in an ideal situation, however, with perfect calibration and no data loss, we should bear in mind that eye-tracking studies still rely on Just and Carpenter's eye-mind assumption (1980: 331) that what the eyes are fixating (the "outside") correlates somehow with what the mind is processing (the "inside"); therefore, "eye movements are a window on the mind, but not necessarily a very clean and fully transparent window" (Jakobsen 2014: 66).

In sum, the possibilities for data exploration in translation process research are immense, and the temptation is to use ever-more complex tools in the data-gathering process. If, however, we want to carry out research without displacing translators from their regular work environment, some problems will inevitably arise, and solutions will have to be found.

I have realised that the "naturalism" of the experimental setting (Séguinot 1996: 76) is not enough to ensure its ecological validity. I encountered major challenges to some of the methods I used, and I have needed to adapt the breadth and depth of my analysis to the availability and quality of the data at my disposal. Reflecting on the way I addressed these challenges, several general principles seem to emerge:

1. Redundancy in the data collection methods is very important, as when one method does not work, there is still an alternative way of recovering data from a different source. The plurality of data-gathering tools means that several can be used at once, not necessarily in a traditional practice of "triangulation" (where each tool gives data from a different perspective) but also as simple compensation: if data are missing or are doubtful in the feed from one tool, they may be replaced or confirmed with data from another.
2. The search for greater accuracy in our measurements may run counter to the criteria of ecological validity. I saw this particularly in my work with eye tracking, where the choice of equipment and the decision install it on the individual computers enhanced validity, but created other problems later in the experiment.
3. Trade-off solutions are sometimes possible, as in my use of eye-tracking software on the researcher's computer only. That is, a simplification of the

experiment environment for the subject implies a far more complicated environment for the researcher.

4. In dealing with these technical and technological problems, the most viable solution is often a robust research design. In this case, I was not able to use areas of interest based on the eye-tracking data to test some of my assumptions, but I was nevertheless able to address the question by looking at the ways the translators solved a set of focus translation problems.

7.4.3. Qualitative data collection methods

Interviews were performed immediately after the last translation task had been finished, similarly to what Hansen (2006; 2008) calls "immediate dialogue". The questions focused on the translators' perception of speed, quality and "comfort" in relation to the different work environments. This qualitative data was later tabulated and compared with the quantitative data obtained with the other data collection methods (section 5.2). This allowed me to confront their perceived performance with the measured performance, and I found that they barely correlate, which is an interesting finding.

The retrospection with replay, on the other hand, was less successful than I had anticipated. The fact of replaying the translation process together with the dynamic eye-tracking data proved very distracting: the translators were rather following their fixations and saccades on screen than actually paying attention to the steps involved in their translation, which were my main goal. I had to ask specific questions and point some tricky problems in the text to receive some still scant feedback.

In the interviews, by asking the question "In which environment did you feel more comfortable?", I initially assumed that "comfortable" (or "at ease", in Spanish: "cómodo") might inversely correlate with typing effort. This proved to be a very naive assumption, as other authors had already pointed at typing as just one component of the effort involved in the translation activity (cf. the "technical" component proposed by Krings 2001 for post-editing effort). Moreover, comfort seems to be related more to long-time experiential factors than to momentary task characteristics. If a similar experiment is reproduced, a question to be asked should be simply "In which environment do you think you typed more?". Alternatively, a different measurement for effort should be used.

A last point to be made regarding the interview data in this study is that not all the translators commented on all the tasks (see Table 15). For a complete comparison between quantitative and qualitative data, a better strategy should be found to elicit answers for the variables in all tasks, while still making sure the answers are not influenced by the researcher's prompts. Be that as it may, I consider the interviews have still provided enough information to draw relevant conclusions about the translators' perceptions.

7.4.4. Choice of translation tool

When planning my experiment and deciding which TM system to use, I considered several options, especially the mainstream ones such as SDL Trados Studio, memoQ, Wordfast Pro and Déjà Vu. I finally chose to use IBM TranslationManager for the reasons explained in section 4.5.1, related to the human and material resources available at the company where the experiment was to be conducted. The graphical user interface of IBM TranslationManager is still based on the old Windows 2000 look and feel, but the tool has nevertheless kept pace with most of the latest developments as far as functionality is concerned, especially its integration of MT. In any case, my focus has not been on the specific tool, but on the usability principles that govern the use of translation memory systems in general, as far as the presentation of translation suggestions and metadata is concerned. In that sense, regardless of the tool used in the experiment, I believe the discussions are of a general nature and can be extended to any other tool working under the same principles.

Having said this, it is worth noting that some discomfort was reflected in several comments gathered during the post-performance interviews because of a pop-up window that appeared every time the translators moved from one segment to the next in the Blind task, saying that the translation of the document was completed. Apparently, the tool is not intended to be used with pre-inserted text, or text is supposed to be pre-inserted only for those segments that are to be skipped (i.e. exact matches). Although the translators just needed to press Enter to dismiss the window, its mere showing up might have had disturbing effects. This also complicated the analysis of the eye-tracking data, as several fixations on a particular area of the screen did not represent fixations on the translation pane located on the same area, but were fixations on this prompt window. Since the eye-tracking data was used in a very restricted way, it was possible to work around this problem by eliminating from the analysis the period during which the pop-up window was displayed. A full eye-tracking analysis of the entire dataset would have made it necessary to discount the fixations on this region from the time the window appeared until the time it disappeared from view, for example by manually activating and deactivating the areas of interest.

In any case, it would be interesting to replicate this study with tools with different window layouts in order to assess whether the way the data are displayed also affects translators' behaviour (besides the mere fact of displaying/not displaying them).

7.4.5. Choice of materials

Even though we took pains to have segments of comparable lengths, real texts usually present a great variation in the number of words per sentence. Although I extracted the source texts from an IBM software manual, which follows controlled language rules, there was still had a great variation of segment lengths (from 7 to 41 words). According to Plitt and Masselot (2010: 12), "20–25 word sentences are probably more likely to be semantically self-contained" and to produce better results for post-editing. Even for human translation, "[a]n optimum throughput appears to be reached for sentences of around 25 words" (ibid.). In my experiment, a compromise had to be made between the ideal length of the segments and the authenticity of the source texts, and the latter ended up receiving higher priority. This factor is not expected to have created a bias in the results, as according to the study just mentioned the irregular lengths can affect both TM and MT suggestions, and the four types of translation suggestions were randomly assigned to the segments.

Another issue related to the selection of source texts refers to the comparability between the texts used for the two main translation tasks. Even if the Flesch-Kincaid tests showed very similar results for both texts, the statistical analysis showed a significant effect of Text on Translation Time, indicating that in practice the texts did not require the same amount of time to translate. This potential effect had been considered and was anticipated by assigning the source texts alternately among the translators and between the tasks, and by including it as a factor in the statistical model.

7.4.6. Data analysis

For the data analysis, I resorted to both quantitative and qualitative methods. The quantitative data coming from keystroke logging and quality assessment were analysed statistically using basic descriptive and inferential statistics (section 5.1) and then with mixed-effects regression models (section 5.4). A great benefit of the mixed-effects models is that they make it easier to investigate simultaneously the effects of several variables and potential confounds on the dependent variables, thus also allowing one to test the validity of the experimental design. An alternative would be the use of ANOVA (analysis of variance) with repeated measures, but this analysis would be more complicated if I wanted to include all the factors that were taken into account in the regression models. Moreover, ANOVA only allows for the inclusion of factors (categorical independent variables) in the analysis, whereas the regression models allowed me to include some covariates (even if they proved to be non-significant).

A major difficulty in the statistical analysis was the need to split the analysis into two steps due to the nature of two of the variables being investigated. Although counter-intuitive at first consideration, the use of the two-step method was the most suitable option to handle the data distribution I had, as corroborated by two professional statisticians I consulted. Eventually, as the analysis progressed, the method provided results that were also relatively straightforward to interpret.

The qualitative data obtained in the interviews and retrospection was not as complete as originally intended (see section 7.4.3) and did not allow me to make a full comparison between the qualitative and quantitative data regarding performance vs. perception (section 5.2). Yet, these data were useful to elicit opinions from the participants about certain aspects of their performances and to check for potential problems in the experimental set-up (section 5.3).

7.4.6.1. Quality assessment

Quality assessment deserves special mention because it remains an unresolved topic in Translation Studies, and I must admit that I have not tackled the issue in any innovative way. The quality assessment method used in this thesis was based on the standard practice in the industry, which mostly relies on error-score systems, with the LISA grid still hovering as a strong reference. Other current initiatives follow along the same lines, as is the case with the newer Dynamic Quality Framework (DQF) proposed by TAUS (2013). While we know that any kind of human revision is always subjective (one can get as many different grades as the number of evaluators), I tried to compensate for this in my analysis by having the two tasks performed by the same translator assessed by the same evaluator, so as to make sure that both tasks were always evaluated according to the same criteria. In addition, each translator was assessed by both evaluators and the level of agreement between the two was very high.

One potential source of errors, however, was the reference translation memory used for creating the translation memories for the experiment, which contained some of the errors that were marked by the reviewers in the translations. If a similar experiment is to be reproduced in the future, it would be advisable to have the reference translation memory reviewed by the same reviewers that will do the quality assessment later, before creating the different types of TM matches. An alternative would be to discount these errors or to analyse them separately. In any case, a simple observation when checking the reviewer's corrections suggests there were just a few occurrences of such errors.

A third problem related to quality assessment concerns an issue brought up by one of the reviewers, who mentioned that the method used for the reviews was very different from their "normal" review method. He referred to the fact that he had to read 10 translations (one by each translator) for the same segment, whereas in a normal situation he would have read only one translation for each segment. I must recognise that despite my concerns with ecological validity, the revision process was also part of the experiment and as such was done under some unnatural conditions.

7.4.7. The human factor

The small number of participants is a common trait of translation process studies, either because it is difficult to find enough people with similar profiles who are willing to participate in an experiment (often for free) or because of the data explosion resulting from the collection methods used in such studies, which increases with the number of participants.

Another difficulty is the great level of inter-subject variation, since the results in these kinds of studies depend on individual differences between the participants, not only on the main variables that we want to investigate. This is true in terms of translation styles, linguistic beliefs and personalities, but it also depends on emotional factors and intuition (Hubscher-Davidson 2009; 2013).

The fact of presenting a near-authentic translation project to translators activates a multitude of mental assumptions and behaviours based on elements such as their previous experiences related to that particular client, the instructions they have received over the years, how they perceive the quality demands from feedback they have obtained on previous jobs, and their personal attitudes towards the type of material. Such things are part and parcel of workplace research, and we must learn to live with them.

7.4.8. Translation instructions

My experiment was strongly "observational", since I controlled its conditions to the minimum extent possible: I was analysing how translators dealt with IBM projects normally, rather than giving explicit instructions on how to configure the tool, how to handle quality, etc. This made it possible to investigate performances that were closer to the real world, but at the same time allowed more variations between the participants than would have happened in a more controlled setting. For example, the translators were left free to choose which key combination to use in the tool to move from one segment to the next. This was responsible for a minor difference in performance between the translators, especially because three women used a keyboard shortcut to move between the segments that skipped exact matches automatically in the Visual task. This difference caused gender and metadata to produce some significant effects (sections 6.2.5 and 6.2.6). If variations such as this are to be avoided in other studies, more explicit instructions should be given.

Likewise, some participants asked several questions before starting the task, while others did not express a single doubt. My answers were usually what was already in the limited instructions I provided (basically: "Do as you would normally do in an IBM project"). However, the fact of asking and receiving an answer might have made some translators feel more confident or even understand better what was expected from them. This short dialogue with (very talkative) P01 before the Visual task illustrates the point:

> P01: In theory, when we translate for IBM, we respect the suggestion in the memory. If the memory gives me a fuzzy match, in principle I consider it is correct. For example, if I detect in the Source of Proposal pane that the only thing that changes is the end, maybe I don't even read the rest of the sentence. In this case, should I behave like this or should I read the fuzzy suggestion a bit, in case there are any changes? I'm asking this because in principle we respect what comes in the memory.
>
> Researcher: But if you see an error in the memory when you're translating for IBM, do you change it?
>
> P01: I do, but I don't pay much attention, I mean, I assume the memory that was sent by IBM should already be... you know... Since we later pass it through Xbench or I do a Validation or use the spell checker in Word, so there are things like errors or typos that I pay less attention to, so I don't waste time in the part that has new words.
>
> Researcher: Do as you would do for IBM, because in fact they are authentic memories from IBM. We have prepared IBM memories.

> P01: OK. I'm asking this because the first one is a case where I see that the entire first chunk of the sentence corresponds to the memory, so I will only pay attention to the last chunk. Because I have this pane activated; maybe other translators don't have it like this.

Finally, the instructions for the Blind task contained the word "revisar", where it should have contained the word "traducir" – like in the Visual task. This is a mistake I made when writing the instructions that could also have influenced the participants' perceptions about the tasks (see sections 4.5.7.3 and 6.3).

7.4.9. Learning curve

The participants' experience in working with the different tasks could improve over time – thus affecting their translation times, typing efforts and error scores –, at least as far as post-editing MT is concerned but also with regard to the pre-translation environment. Therefore, the data I gathered might be representative of performance at the beginning of a learning curve. This is especially important as task familiarity was mentioned more than once as a prominent factor by many participants in the interviews. One solution would be to train translators for some period and measure their performance after some time. Within the framework of my doctoral research, it was not possible to carry out a longitudinal study, and the compromise solution I envisaged was to try to find participants with comparable levels of post-editing experience.

7.5. Avenues for future research

The current study allowed me to analyse how translation metadata can help translators deal with translation suggestions when only one suggestion is presented for each segment. Since translating with CAT tools usually involves a dual process of selection plus repairing of suggestions, it would be interesting to complement the current study with a follow-up experiment including multiple suggestions, to investigate how metadata can also help translators choose between different suggestions.

Likewise, the material conditions of my experiment did not allow me to isolate translation metadata as the main independent variable, since "presentation mode" (pre-insertion in the Blind task vs. "dynamic" insertion in the Visual task) was also playing a role (see sections 4.5.7.1 and 6.1.7). The experiment could be improved by isolating the "presentation mode" and "metadata" factors, but the best way of doing this is not obvious, especially if we want to compare the use of TM and MT in the same tool. For example, translation memory systems like the one I used always indicate the type of suggestion whenever a suggestion is available. If I chose to offer dynamic insertion in both tasks, an option for hiding the metadata in the Blind task would be to flag all suggestions as exact matches or as machine translations. However, even if translators were informed beforehand that the type of suggestion presented by the tool in that task was just "for the record", they could still be biased by the false metadata. If, on the other hand, I decided to offer pre-insertion in both tasks, the resulting environment would be different from the one normally used for TM translation. The solution is not simple.

Another topic that could be explored further is how MT suggestions are handled by the translators according to the quality of the individual suggestions. The translation suggestions presented in my experiment included TM matches at three quality levels along with MT suggestions of undistinguished quality levels, mostly of very high quality. It would be interesting to include MT suggestions of different quality levels and classify them (e.g. using BLEU scores) so that they could be compared against the three levels of TM matches.

The current study also raises a question about the potential effects of adding metadata for MT suggestions: just as TM systems display the fuzzy match level and textual differences for TM suggestions, what would be the impact of presenting information about the degree of confidence or areas of uncertainty for MT suggestions? Some implementations based on MT quality estimation have been proposed in experimental tools such as PET (Aziz et al. 2012), MateCat (Federico et al. 2014) and CASMACAT (2013), and in commercial tools such as Asia Online's Language Studio.[20] It is expected that MT metadata can help post-editors in their decision-making processes, but there is still not enough empirical evidence that corroborates this assumption.

These possibilities bring about additional topics for discussion. In an environment that contains TM and MT with metadata, what would be the real differences between repairing TM suggestions and MT suggestions? Would it still make sense to differentiate between "translation" and "post-editing", a term that has traditionally been used to refer to the activity of repairing an MT suggestion as a starting point to produce translations? Since the general tendency seems to be towards environments that combine both TM and MT, it is becoming necessary to either broaden the definition of post-editing to include those mixed scenarios or to drop the use of this term in favour of just "translation", as

[20] http://www.languagestudio.com/LanguageStudioDesktop.aspx#Pro

virtually no translation happens without the help of technology these days, and most of translating is based on previous translations (either coming from TM or MT).

Moving beyond the scope of the experiment and building on related research that has been published in recent years, it seems necessary to take into account different translator styles when analysing the results of studies on translation technologies, as some behaviours cannot be generalised for all translators or for the translation process in general (Hubscher-Davidson 2009). In this sense, it would be interesting to look into how specific features in the translation tools (such as metadata) affect translators with different styles. The CASMACAT project seems to be addressing the issue of different translator styles by analysing the changes in performance metrics under different translation modes (CASMACAT 2013).

Another promising field of research to advance the existing knowledge on translation processes is that of Human-Computer Interaction, as advocated by O'Brien (2012). Concepts such as "cognitive ergonomics" could help us understand usability aspects and avoid "cognitive frictions" (op. cit.: 116) between the way the tools present information (segmentation of source/target texts, metadata, etc.) and the way translators expect that information to be presented. I support O'Brien when she foresees possibilities for increased collaboration between the tool users (translators) and tool developers. Among other benefits, collaborative research in this area could bring more flexibility to the design of tools and improve the interaction possibilities between translators and the technologies that support their work.

Also focusing on the concept of ergonomics (cognitive, physical and organisational), Ehrensberger-Dow and Massey (2014) have conducted research on translators' habits at the workplace, with a view to identifying those elements that can contribute to mental and physical stress among translators. This relates to some of the topics I have mentioned in passing throughout this thesis, such as job satisfaction and the feeling of comfort. Ehrensberger-Dow and Massey call for more research on "how workplace ergonomics can influence translation performance" (op. cit.: 65) and have announced "experiments with different user interface settings in a usability lab" (op. cit.: 80). The results of such studies are also expected to contribute to the understanding of how certain aspects of the tools and workflows affect translators cognitively.

Research in the field also needs to address some of the sociological and humanistic aspects of technology. MT is making it possible to have more translations for free or done by non-professionals, and source-language knowledge might become less important.

Translators need to keep up with the latest tools and learn about MT post-editing, and they must be prepared to work in collaboration not only with their "traditional" partners, such as project managers and terminologists, but also with area experts and even final users, who are gradually becoming involved in the translation workflow. Future research should look further into how these technological changes are affecting translators in several ways (see Garcia 2007; Pym 2012; Temizöz 2013).

The rapid changes in the available technologies also require changes in the training of translation professionals, both for those entering the market for the first time and for those seeking opportunities of lifelong learning. Understanding the interaction between translators and tools in specific tasks – such as "post-editing" or repairing suggestions from translation memories and machine translation – can also help in the design of curricula that are more in line with market needs.

Finally, going back to the issues related to the translator's work environment, several questions remain to be answered with respect to translation metadata. Is it more effective to indicate a TM match type through colour codes, letters, or both? Is it better to have metadata displayed close to the editing area? And going beyond these particularities, how could translation tools be made more intuitive and easier to use, in this era of touch screens and voice commands? Is dictation software about to experience the same boom as machine translation has in the past decade? The future seems promising when thinking about the possibilities of making translation technology not only more productive but also more enjoyable and comfortable to work with.

7.6. Final remarks

This study has looked into the strategies used by professional translators when interacting with a translation tool in a real-world situation. It aimed at contributing to a better understanding of the translation process in terms of how metadata is used. My results have indicated that translation metadata improve productivity for certain types of translation suggestions, especially exact matches and fuzzy matches in the 85%–99% range. The study has also determined that translation metadata do not affect translation quality significantly.

According to the translators' testimonials and the session recordings, even if metadata are not consulted very frequently, it is important that they be there when needed, as this increases translators' confidence in how to repair the suggestion. The presence of metadata was also associated with an increased feeling of task satisfaction by most of the respondents.

As an additional finding, translation times did not correlate with typing effort when metadata are present, which suggests that development efforts for TM tools should focus not on finding aids to help in the typing process (e.g. auto-suggest features) but rather on finding ways of speeding up the decision-making process (finding alternatives, choosing among them and knowing how to repair them).

I believe that more research is needed to identify the pieces of metadata that are most useful to the translation process in terms of both productivity and task satisfaction. Not all tools need present metadata the same way or have similar interfaces, as interface differences are important to account for different translator styles. This calls for increased collaboration between tool developers and tool users, taking on board the developments in other fields such as those of usability and ergonomics.

References

Allen, Jeffrey H. 2003. "Post-editing". In *Computers and Translation: A translator's guide*, Harold L. Somers (ed.) Amsterdam and Philadelphia: John Benjamins. 297–317.

Allen, Jeffrey H. 2005. "What is Post-editing?". *Translation Automation Newsletter* (4): 1–5.

Alves, Fabio (ed.) 2003. *Triangulating Translation: Perspectives in process oriented research*. Amsterdam and Philadelphia: John Benjamins.

Anastasiou, Dimitra, and Lucía Morado Vázquez. 2010. "Localisation standards and metadata". In *Metadata and Semantic Research* [Communications in Computer and Information Science 108], Salvador Sánchez-Alonso and Ioannis N. Athanasiadis (eds). Berlin and Heidelberg: Springer Berlin Heidelberg. 255–274.

Angelone, Erik. 2010. "Uncertainty, uncertainty management and metacognitive problem solving in the translation task". In *Translation and Cognition* [American Translators Association Scholarly Monograph Series 15], Gregory M. Shreve and Erik Angelone (eds). Amsterdam and Philadelphia: John Benjamins. 17–40.

Asadi, Paula, and Candace Séguinot. 2005. "Shortcuts, strategies and general patterns in a process study of nine professionals". *Meta : Journal des traducteurs / Meta: Translators' Journal* 50 (2): 522.

Austermühl, Frank. 2001. *Electronic Tools for Translators* [Translation Practices Explained 2]. Manchester: St. Jerome.

Autodesk. 2011. *Translation and Post-Editing Productivity*. http://langtech.autodesk.com/productivity.html. Accessed October 2014.

Aziz, Wilker, Sheila Sousa and Lucia Specia. 2012. "PET: A tool for post-editing and assessing machine translation". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC]. Istanbul, Turkey. 3982–3987.

Balling, Laura W. 2008. "A brief introduction to regression designs and mixed-effects modelling by a recent convert". In *Looking at Eyes: Eye-tracking studies of reading and translation processing* [Copenhagen Studies in Language 36], Susanne Göpferich, Arnt L. Jakobsen and Inger M. Mees (eds). Frederiksberg: Samfundslitteratur. 175–192.

Bernardini, Silvia. 1999. "Using think-aloud protocols to investigate the translation process: Methodological aspects" [RCEAL Working papers in English and Applied Linguistics 6], N. W. John (ed.) Cambridge: University of Cambridge. 179–199.

Bernth, Arendse. 2006. "EasyEnglishAnalyzer: Taking controlled language from sentence to discourse level". Paper presented at the 5th International Workshop on Controlled Language Applications (CLAW 2006).

Biau Gil, José Ramón. 2005. *Flying blind: Translation interfaces and non-verbal elements in hypermedia texts*. Master's thesis. Tarragona: Universitat Rovira i Virgili.

Bowling, Nathan A., Kevin J. Eschleman and Qiang Wang. 2010. "A meta-analytic examination of the relationship between job satisfaction and subjective well-being". *Journal of Occupational and Organizational Psychology* 83 (4): 915–934.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer and P. Roossin. 1988. "A statistical approach to language translation". In *Proceedings of the 12th Conference on Computational Linguistics* [1], Denes Vargha (ed.) Budapest, Hungary. 71–76.

Bruckner, Christine, and Mirko Plitt. 2001. "Evaluating the operational benefit of using machine translation output as translation memory input". In *Machine Translation in the Information Age (Proceedings of the 8th MT Summit organised by the European Association for Machine Translation - MT evaluation workshop),* Bente Maegaard (ed.). Santiago de Compostela, Spain, 18-22 September.

Buchweitz, Augusto, and Fabio Alves. 2006. "Cognitive adaptation in translation: An interface between language direction, time, and recursiveness in target text production". *Letras de Hoje* 41 (2): 241–272.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt and Arnt L. Jakobsen. 2011. "The process of post-editing: A pilot study". In *Proceedings of the 8th International NLPCS Workshop. Special theme: Human-Machine Interaction in Translation* [Copenhagen Studies in Language 41], Bernadette Sharp, Michael Zock, Michael Carl and Arnt L. Jakobsen (eds). Frederiksberg: Samfundslitteratur. 131–142.

Carl, Michael, and Silvia Hansen. 1999. "Linking translation memories with example-based machine translation". In *MT in the Great Translation Era (Proceedings of*

Machine Translation Summit VII '99). Kent Ridge Digital Labs, Singapore, September 13-17. 617–624.

Carl, Michael, Martin Kay and Kristian T. Jensen. 2010. "Long distance revisions in drafting and post-editing". Paper presented at CICLing-2010, Iaşi, Romania.

CASMACAT. 2013. *Public Second Year Report*. http://www.casmacat.eu/index.php?n=Main.SecondYear. Accessed October 2014.

CASMACAT. 2014. *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*. http://www.casmacat.eu/. Accessed September 2014.

Christensen, Tina P. 2011. "Studies on the mental processes in translation memory-assisted translation: The state of the art". *trans-kom. Zeitschrift für Translationswissenschaft und Fachkommunikation* 4 (2): 137–160.

Christensen, Tina P., and Anne Schjoldager. 2010. "Translation-memory (TM) research: What do we know and how do we know it?". *Hermes – Journal of Language and Communication Studies* 44: 89–101.

Colominas, Carme. 2008. "Towards chunk-based translation memories". *Babel* 54 (4): 343–354.

Danks, Joseph H., Gregory M. Shreve, Stephen B. Fountain and M. McBeath (eds). 1997. *Cognitive Processes in Translation and Interpreting* [Applied Psychology 3]. Thousand Oaks, CA: Sage Publications.

de Almeida, Giselle. 2013. *Translating the post-editor: An investigation of post-editing changes and correlations with professional experience across two Romance languages*. Doctoral thesis. Dublin: Dublin City University.

Dillon, Sarah, and Janet Fraser. 2006. "Translators and TM: An investigation of translators' perceptions of translation memory adoption". *Machine Translation* 20 (2): 67–79.

Dimitrova, Birgitta E. 2005. *Expertise and Explicitation in the Translation Process* [Benjamins Translation Library 64]. Amsterdam: John Benjamins.

Doherty, Stephen, and Joss Moorkens. 2013. "Investigating the experience of translation technology labs: Pedagogical implications". *The Journal of Specialised Translation* (19): 122–136.

Dragsted, Barbara. 2004. *Segmentation in translation and translation memory systems: An empirical investigation of cognitive segmentation and effects of integrating a*

TM system into the translation process. Doctoral thesis. Frederiksberg: Copenhagen Business School.

Dragsted, Barbara. 2005. “Segmentation in translation: Differences across levels of expertise and difficulty”. *Target* 17 (1): 49–70.

Dragsted, Barbara. 2012. “Indicators of difficulty in translation — Correlating product and process data”. *Across Languages and Cultures* 13 (1): 81–98.

Ehrensberger-Dow, Maureen. 2014. “Challenges of translation process research at the workplace”. In *Minding translation / Con la traducción en mente*, Ricardo Muñoz Martín (ed.). San Vicente del Raspeig: Publicaciones de la Universidad de Alicante. 355–383.

Ehrensberger-Dow, Maureen, and Gary Massey. 2014. “Cognitive ergonomic issues in professional translation”. In *The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science*, John W. Schwieter and Aline Ferreira (eds). Newcastle upon Tyne: Cambridge Scholars Publishing. 58–86.

Enríquez Raído, Vanessa. 2011. *Investigating the Web search behaviors of translation students: An exploratory and multiple-case study*. Doctoral thesis. Barcelona: Universitat Ramon Llull.

Ericsson, K. A., and Herbert A. Simon. 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Ericsson, K. A., and Herbert A. Simon. 1998. “How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking”. *Mind, Culture, and Activity* 5 (3): 178–186.

Esselink, Bert. 2000. *A Practical Guide to Localization* [Language International World Directory 4]. Amsterdam and Philadelphia: John Benjamins.

Federico, Marcello, Nicola Bertoldi, M. Cettolo, M. Negri, M. Turchi, M. Trombetti, A. Cattelan, A. Farina, D. Lupinetti, A. Martines, A. Massidda, H. Schwenk, L. Barrault, F. Blain, Philipp Koehn, C. Buck and U. Germann. 2014. “The Matecat tool”. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. 129–132.

Ferreira, Aline, and John W. Schwieter (eds). Forthcoming. *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. John Benjamins.

Garcia, Ignacio. 2007. “Power shifts in web-based translation memory”. *Machine Translation* 21 (1): 55–68.

Garcia, Ignacio. 2010. "Is machine translation ready yet?". *Target* 22 (1): 7–21.

Göpferich, Susanne, Fábio Alves and Inger M. Mees (eds). 2009. *New Approaches in Translation Process Research* [Copenhagen Studies in Language 39]. Frederiksberg: Samfundslitteratur.

Göpferich, Susanne, Arnt L. Jakobsen and Inger M. Mees (eds). 2008. *Looking at Eyes: Eye-tracking studies of reading and translation processing* [Copenhagen Studies in Language 36]. Frederiksberg: Samfundslitteratur.

Göpferich, Susanne, Arnt L. Jakobsen and Inger M. Mees (eds). 2009. *Behind the Mind: Methods, models and results in translation process research* [Copenhagen Studies in Language 37]. Frederiksberg: Samfundslitteratur.

Gorden, Raymond L. 1992. *Basic Interviewing Skills*. Itasca, IL: F.E. Peacock.

Gouadec, Daniel. 2007. *Translation as a Profession* [Benjamins Translation Library 73]. Amsterdam and Philadelphia: John Benjamins.

Green, Spence, Jeffrey Heer and Christopher D. Manning. 2013. "The efficacy of human post-editing for language translation". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Wendy E. Mackay, Stephen Brewster and Susanne Bødker (eds). Paris, France. 439–448.

Guerberof Arenas, Ana. 2009. "Productivity and quality in the post-editing of outputs from translation memories and machine translation". *Localisation Focus - The International Journal of Localisation* 7 (1): 11–21.

Guerberof Arenas, Ana. 2012. *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. Doctoral thesis. Tarragona: Universitat Rovira i Virgili.

Guerberof Arenas, Ana. 2013. "What do professional translators think about post-editing?". *The Journal of Specialised Translation* (19): 75–95.

Guerra Martínez, Lorena. 2003. *Human translation versus machine translation and full post-editing of raw machine translation output*. Master's thesis. Dublin: Dublin City University.

Hansen, Gyde. 2005. "Experience and emotion in empirical translation research with think-aloud and retrospection". *Meta: Journal des traducteurs / Meta: Translators' Journal* 50 (2): 511–521.

Hansen, Gyde. 2006. "Retrospection methods in translator training and translation research". *The Journal of Specialised Translation* 5: 2–41.

Hansen, Gyde. 2008. "The dialogue in translation process research". In *Translation and Cultural Diversity: Selected Proceedings of the XVIII FIT World Congress 2008 (XVIII FIT World Congress 2008)*. Shanghai, China: Foreign Languages Press.

Hartmann, Nicholas. 2010. "Real voices: What translators do and why we need to keep doing it". Keynote presentation at the 9th Conference of the Association for Machine Translation in the Americas (AMTA 2010). Denver, CO, 31 October - 4 November.

He, Yifan, Yanjun Ma, Johann Roturier, Andy Way and Josef van Genabith. 2010. "Improving the post-editing experience using translation recommendation: A user study". Presentation at the 9th Conference of the Association for Machine Translation in the Americas (AMTA 2010). Denver, CO, 31 October - 4 November.

House, Juliane. 2000. "Consciousness and the strategic use of aids in translation". In *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on empirical research*, Sonja Tirkkonen-Condit and Riitta Jääskeläinen (eds). Amsterdam and Philadelphia: John Benjamins. 149–162.

Huang, Harry J. 2011. "Intermediality and human vs. machine translation". *CLCWeb: Comparative Literature and Culture* 13 (3). 1–11.

Hubscher-Davidson, Séverine E. 2013. "The role of intuition in the translation process: A case study". *Translation and Interpreting Studies* 8 (2): 211–232.

Hubscher-Davidson, Séverine E. 2009. "Personal diversity and diverse personalities in translation: A study of individual differences". *Perspectives: Studies in Translatology* 17 (3): 175–192.

Hutchins, W. J., and Harold L. Somers. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Hvelplund, Kristian T. 2011. *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Doctoral thesis. Frederiksberg: Copenhagen Business School.

Hvelplund, Kristian T. 2014. "Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data". In *Minding Translation / Con la traducción en mente*, Ricardo Muñoz Martín (ed.) San Vicente del Raspeig: Publicaciones de la Universidad de Alicante. 201–224.

Jääskeläinen, Riitta. 1993. "Investigating translation strategies". In *Recent Trends in Empirical Translation Research* [Kielitieteellisiä tutkimuksia, Studies in

Languages 28], Sonja Tirkkonen-Condit and John Laffling (eds). Joensuu: Joensuu University. 99–120.

Jääskeläinen, Riitta. 2002. "Think-aloud protocol studies into translation: An annotated bibliography". *Target* 14 (1): 107–136.

Jakobsen, Arnt L. 2002. "Translation drafting by professional translators and by translation students". *Copenhagen Studies in Language* (27): 191–204.

Jakobsen, Arnt L. 2003. "Effects of think aloud protocols on translation speed, revision and segmentation". In *Triangulating Translation: Perspectives in process oriented research*, Fabio Alves (ed.) Amsterdam and Philadelphia: John Benjamins. 69–95.

Jakobsen, Arnt L. 2006. "Research methods in translation – Translog". In *Computer keystroke logging and writing: Methods and applications* [Studies in Writing 18], Eva Lindgren and Sullivan, Kirk P. H. (eds). Amsterdam: Elsevier. 95–105.

Jakobsen, Arnt L. 2014. "Theoretical and methodological aspects of translation process research". Presentation at the Fourth International PhD course in Translation Process Research (TPR 2014). Copenhagen, Denmark, July 7 – 11.

Jakobsen, Arnt L., and Kristian T. Jensen. 2008. "Eye movement behaviour across four different types of reading task". In *Looking at Eyes: Eye-tracking studies of reading and translation processing* [Copenhagen Studies in Language 36], Susanne Göpferich, Arnt L. Jakobsen and Inger M. Mees (eds). Frederiksberg: Samfundslitteratur. 103–124.

Johnson, R. B., A. J. Onwuegbuzie and L. A. Turner. 2007. "Toward a definition of mixed methods research". *Journal of Mixed Methods Research* 1 (2): 112–133.

Just, Marcel A., and Patricia A. Carpenter. 1980. "A theory of reading: From eye fixations to comprehension". *Psychological Review* 87 (4): 329–354.

Karamanis, Nikiforos, Saturnino Luz and Gavin Doherty. 2011. "Translation practice in the workplace: contextual analysis and implications for machine translation". *Machine Translation* 25 (1): 35–52.

Kay, Martin. 1980. *The proper place of men and machines in language translation*. Palo Alto, CA: Xerox Palo Alto Research Center.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. "Moses: Open source toolkit for statistical machine translation". In *Proceedings* *of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. 177–180. Prague, Czech Republic: Association for Computational Linguistics.

Koponen, Maarit, Wilker Aziz, Luciana Ramos and Lucia Specia. 2012. “Post-editing time as a measure of cognitive effort”. In *Proceedings of the Workshop on Post-Editing Technology and Practice* [WPTP]. 11–20. San Diego, CA, USA, October 28 – November 1.

Krings, Hans P. 1986. *Was in den Köpfen von Übersetzern vorgeht: Eine empirische Untersuchung zur Struktur des Ubersetzungsprozesses an fortgeschrittenen Französischlernern* [Tübinger Beiträge zur Linguistik 291]. Tübingen: Gunter Narr.

Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent, OH: Kent State University Press.

Kussmaul, Paul, and Sonja Tirkkonen-Condit. 1995. “Think-aloud protocol analysis in Translation Studies”. *TTR : traduction, terminologie, rédaction* 8 (1): 177–199.

Labov, William. 1972. *Sociolinguistic patterns* [Conduct and Communication 4]. Philadelphia, PA: University of Pennsylvania Press.

Lagoudaki, Elina. 2006. *Translation Memory systems: Enlightening users’ perspective (Translation Memories Survey 2006)*. Imperial College London.

Lagoudaki, Elina. 2008. *Expanding the possibilities of translation memory systems: From the translator’s wishlist to the developer’s design*. Doctoral thesis. London: Imperial College London.

Lange, Carmen A., and Winfield S. Bennett. 2000. “Combining machine translation with translation memory at Baan”. In *Translating into Success: Cutting-edge strategies for going multilingual in a global age* [American Translators Association Scholarly Monograph Series 11], Robert C. Sprung (ed.). Amsterdam and Philadelphia: John Benjamins. 203–218.

Lee, Jason, and Posen Liao. 2011. “A comparative study of human translation and machine translation with post-editing”. *Compilation and Translation Review* 4 (2): 105–149.

Leijten, Marielle, and Luuk van Waes. 2013. “Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes”. *Written Communication* 30 (3): 358–392.

Liu, Fung-Ming C. 2011. *A quantitative and qualitative inquiry into translators' visibility and job-related happiness: The case of Greater China*. Doctoral thesis. Tarragona: Universitat Rovira i Virgili.

Lörscher, Wolfgang. 1991. *Translation Performance, Translation Process and Translation Strategies: A psycholinguistic investigation*. Tübingen: Gunter Narr.

Martín-Mor, Adrià. 2011. *La interferència lingüística en entorns de Traducció Assistida per Ordinador: Recerca empíricoexperimental*. Doctoral thesis. Bellaterra: Universitat Autònoma de Barcelona.

McBride, Cheryl. 2009. *Translation memory system: An analysis of translator's attitudes and opinions*. Master's thesis. Ottawa: University of Ottawa.

McKay, Corinne. 2011. "Webinar question: How many words per day?". *Thoughts on Translation*. http://thoughtsontranslation.com/2011/01/20/webinar-question-how-many-words-per-day. Accessed October 2014.

Melby, Alan K. 1982. "Multi-level translation aids in a disbributed system". In *Proceedings of the Ninth International Conference on Computational Linguistics (COLING82)* [Linguistic Series 47], Jan Horecky (ed.) 215–220. Amsterdam: North-Holland Publishing Company.

Melby, Alan K. 2013. *Interview to Jost Zetzsche*. http://www.internationalwriters.com/BigWave/BigWaveAKM.htm. Accessed October 2014.

Moorkens, Joss. 2012. *Measuring consistency in translation memories: A mixed-methods case study*. Doctoral thesis. Dublin: Dublin City University.

Morado Vázquez, Lucía. 2012. *An empirical study on the influence of translation suggestions' provenance metadata*. Doctoral thesis. Limerick: University of Limerick.

Morado Vázquez, Lucía, and Jesús Torres del Rey. 2011. "The relevance of metadata during the localisation process – an experiment". Paper presented at the First International T3L Conference: Tradumatica, Translation Technologies and Localization). Bellaterra: Universitat Autònoma de Barcelona.

Moran, John. 2014. "The impact of touch typing on words per day productivty" (sic). *ProZ.com forum*. http://www.proz.com/forum/translation_theory_and_practice/262922-the_impact_of_touch_typing_on_words_per_day_productivty.html. Accessed October 2014.

Mossop, Brian. 2001. *Revising and Editing for Translators* [Translation Practices Explained 3]. Manchester: St. Jerome.

Mossop, Brian. 2007. "Empirical studies of revision: A review". *The Journal of Specialised Translation* (8). 5–20.

Muñoz Martín, Ricardo (ed.) 2014. *Minding Translation / Con la traducción en mente*. San Vicente del Raspeig: Publicaciones de la Universidad de Alicante.

Nagao, Makoto. 1984. "A framework of a mechanical translation between Japanese and English by analogy principle". In *Artificial and Human Intelligence: Edited review papers presented at the International NATO Symposium held in Lyon, France, October, 1981*, Alick Elithorn and Ranan Banerji (eds). Amsterdam: North-Holland Publishing Company. 173–180.

Nyberg, Eric, Teruko Mitamura and Willem-Olaf Huijsen. 2003. "Controlled language for authoring and translation". In *Computers and Translation: A translator's guide*, Harold L. Somers (ed.) Amsterdam and Philadelphia: John Benjamins. 245–281.

O'Brien, Sharon. 2002. "Teaching post-editing: A proposal for course content". In *Teaching Machine Translation (Proceedings of the 6th EAMT Workshop), Centre for Computational Linguistics, UMIST, Manchester, England, November 14-15, 2002*. 99–106.

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

O'Brien, Sharon. 2006b. "Pauses as indicators of cognitive effort in post-editing machine translation output". *Across Languages and Cultures* 7 (1): 1–21.

O'Brien, Sharon. 2009a. "Eye tracking in translation process research: Methodological challenges and solutions". In *Methodology, Technology and Innovation in Translation Process Research: A tribute to Arnt Lykke Jakobsen* [Copenhagen Studies in Language 38], Inger Mees, Fabio Alves and Susanne Göpferich (eds). Frederiksberg: Samfundslitteratur. 251–266.

O'Brien, Sharon. 2009b. "Translation memory interfaces and attention shifts". *Presentation at Eye-to-IT conference, April 2009.*

O'Brien, Sharon. 2010. "Controlled language and readability". *Translation and Cognition* (15): 143–168.

O'Brien, Sharon (ed.) 2011. *Cognitive Explorations of Translation* [Continuum Studies in Translation]. New York, NY: Continuum.

O'Brien, Sharon. 2012. "Translation as human-computer interaction". *Translation Spaces* 1 (1): 101–122.

O'Brien, Sharon, and Joss Moorkens. 2014. "Towards intelligent post-editing interfaces". In *Proceedings of the XXth FIT World Congress,* W. Baur, B. Eichner, S. Kalina, N. Kessler, F. Mayer and J. Orsted (eds). 131–137. Berlin, Germany, August 4 – 6.

O'Hagan, Minako. 2013. "The impact of new technologies on Translation Studies: A technological turn?". In *The Routledge Handbook of Translation Studies* [Routledge Handbooks in Applied Linguistics], Carmen Millán and Francesca Bartrina (eds). Hoboken: Taylor and Francis. 521–536.

PACTE. 2005. "Investigating translation competence: Conceptual and methodological issues". *Meta : Journal des traducteurs / Meta: Translators' Journal* 50 (2): 609–619.

Piróth, Attila. 2011. "Translation automation survey among translators". http://www.pirothattila.com/APiroth_MT-Survey.pdf. Accessed November 2014.

Plitt, Mirko, and François Masselot. 2010. "A productivity test of statistical machine translation post-editing in a typical localisation context". *The Prague Bulletin of Mathematical Linguistics* 93: 7–16.

Pym, Anthony. 2004. *The Moving Text: Localization, translation, and distribution*. Amsterdam and Philadelphia: John Benjamins.

Pym, Anthony. 2010. *Exploring Translation Theories*. London and New York: Routledge.

Pym, Anthony. 2011a. "Translation research terms: A tentative glossary for moments of perplexity and dispute". In *Translation Research Projects 3,* Anthony Pym (ed.) Tarragona: Intercultural Studies Group. 75–110.

Pym, Anthony. 2011b. "What technology does to translating". *Translation & Interpreting* 3 (1): 1–9.

Pym, Anthony. 2012. "Democratizing translation technologies: The role of humanistic research". In *Proceedings of the Language and Translation Automation Conference,* Valeria Cannavina and Anna Fellet (eds). 14–29. Rome: The Big Wave.

Rydning, Antin F. 2002. "Brief introduction to the methodology of Translog and Think Aloud Protocols (TAPs)". [unpublished]

Schwieter, John W., and Aline Ferreira (eds). 2014. *The Development of Translation Competence: Theories and methodologies from Psycholinguistics and Cognitive Science*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Séguinot, Candace. 1996. “Some thoughts about think-aloud protocols”. *Target* 8 (1): 75–95.

Séguinot, Candace. 2000. “Management issues in the translation process”. In *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on empirical research*, Sonja Tirkkonen-Condit and Riitta Jääskeläinen (eds). Amsterdam and Philadelphia: John Benjamins. 143–148.

Sharmin, Selina, Oleg Spakov, Kari-Jouko Räihä and Arnt L. Jakobsen. 2008. “Effects of time pressure and text complexity on translators’ fixations”. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, ETRA’08*. 123–126. ACM.

Shreve, Gregory M., and Erik Angelone (eds). 2010. *Translation and Cognition* [American Translators Association Scholarly Monograph Series 15]. Amsterdam and Philadelphia: John Benjamins.

Skadiņš, R., M. Puriņš, I. Skadiņa and A. Vasiļjevs. 2011. “Evaluation of SMT in localization to under-resourced inflected language”. In *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT 2011), Leuven, Belgium, 30-31 May 2011*, Mikel L. Forcada, Heidi Depraetere and Vincent Vandeghinste (eds). 35–40.

Somers, Harold L. (ed.) 2003. *Computers and Translation: A translator’s guide*. Amsterdam and Philadelphia: John Benjamins.

TAUS. 2013. *Quality Evaluation Using an Error Typology Approach*. https://evaluation.taus.net/resources/error-typology-guidelines. Accessed October 2014.

Teixeira, Carlos S. C. 2011. “Knowledge of provenance and its effects on translation performance in an integrated TM/MT environment”. In *Proceedings of the 8th International NLPCS Workshop. Special theme: Human-Machine Interaction in Translation* [Copenhagen Studies in Language 41], Bernadette Sharp, Michael Zock, Michael Carl and Arnt L. Jakobsen (eds). 107–118. Frederiksberg: Samfundslitteratur.

Teixeira, Carlos S. C. 2014a. “Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories”. In

Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3), Vancouver, BC, Canada, October 22 – 26, 2014, Sharon O'Brien, Michel Simard and Lucia Specia (eds). 45–59.

Teixeira, Carlos S. C. 2014b. "The handling of translation metadata in translation tools". In *Post-Editing of Machine Translation: Processes and applications*, Sharon O'Brien, Laura Balling, Michael Carl, Michel Simard and Lucia Specia (eds). Newcastle upon Tyne: Cambridge Scholars Publishing. 109–125.

Temizöz, Özlem. 2013. *Postediting machine translation output and its revision: Subject-matter experts versus professional translators*. Doctoral thesis. Tarragona: Universitat Rovira i Virgili.

Tobii Technology. 2008. "User Manual: Tobii X60 & X120 Eye Trackers". http://www.tobii.com/Global/Analysis/Downloads/User_Manuals_and_Guides/Tobii_X60_X120_UserManual.pdf. Accessed October 2014.

Tobii Technology. 2010. "Tobii Eye Tracking: An introduction to eye tracking and Tobii eye trackers". White paper. http://www.tobii.com/Global/Analysis/Training/WhitePapers/Tobii_EyeTracking_Introduction_WhitePaper.pdf?epslanguage=en. Accessed October 2014.

Underwood, Nancy, Bartolomé Mesa-Lao, Mercedes García Martínez, Michael Carl, Vicent Alabau, Jesús González-Rubio, Luis A. Leiva, Germán Sanchis-Trilles, Daniel Ortíz-Martínez and Francisco Casacuberta. 2014. "Evaluating the effects of interactivity in a post-editing workbench". In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*. 553–559.

van der Meer, Jaap. 2014. "'Perfect storm' conditions for machine translation". *TAUS Review of language business and technology* (1). https://www.taus.net/taus-review#october-2014. Accessed November 2014.

Wallis, Julian. 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*. Master's thesis. Ottawa: University of Ottawa.

Way, Andy. 2009. "A critique of Statistical Machine Translation". *Linguistica Antverpiensia* (8): 17–41.

Webb, Lynn E. 1998. *Advantages and disadvantages of translation memory: A cost-benefit analysis*. Master's thesis. Monterey: Monterey Institute of International Studies.

Yamada, Masaru. 2011. *Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process*. Doctoral thesis. Rikkyo University.

Zetzsche, Jost. 2014. *The Translator's Tool Box: A computer primer for translators*. [electronic book]: International Writers' Group, LLC.

Appendices

Appendix 1 – Source text used in the Copy task

Si no puede iniciar una sesión satisfactoriamente en el servidor de portal para iniciar la sesión de trabajo de Tivoli Enterprise Portal, revise los síntomas y las acciones de corrección para solucionar el problema.

En la tabla siguiente encontrará las resoluciones para problemas de inicio de sesión en Tivoli Enterprise Portal Server.

Appendix 2 – Source text used in the Scratch task

When installing on AIX® systems, security policies for newly created users auto-expire the password after the first use and require you to set a new (or same) password as a permanent password. The Tivoli® Enterprise Portal Server configuration interface allows you to create a new user ID for the portal server and warehouse database, but using the interface always fails because the user password is not set and is expired. You must ssh/telnet into the same server, using the target user ID, and set the password appropriately.

A **database server** maintains the databases and processes requests from the client to extract data from or to update the database. An **application server** provides additional business-support processing for the clients.

Appendix 3 – SourceText31 (used in the Visual and Blind tasks)

Text_31

Introduction to troubleshooting

You might not always be able to solve a problem yourself after determining its cause. For example, a performance problem might be caused by a limitation of your hardware. If you are unable to solve a problem on your own, contact IBM Software Support for a solution. See Logs and data collection for troubleshooting for information on the types of data to collect before contacting Support.

Trace data capture transient information about the current operating environment when a component or application fails to operate as designed. IBM Software Support personnel use the captured trace information to determine the source of an error or unexpected condition. See Trace logging for more information about tracing.

You can subscribe to e-mail notification about product tips and newly published fixes through the Support portal. In the Support portal, you can specify the products for which you want to receive notifications; choose from flashes, downloads, and technotes; and set up to receive email updates.

1. Open the http://ibm.com website and select Support & downloads > Technical support. You can also launch an IBM® support website.
2. In the Quick start page or Support home, click Sign in to sign in or to register if you have not yet registered.
3. In the Notifications area of Support home, click Manage all my subscriptions.
4. In the Subscribe and My defaults tabs, select a product family and continue setting your preferences to specify the information you want in your emails.
5. If you have not yet added an email address to your profile, click My IBM > Profile > Edit and add it to your personal information.

Glossary

Advanced Encryption Standard

An encryption algorithm for securing sensitive but unclassified material designed by the National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce. AES is intended to be a more robust replacement for the **Data Encryption Standard**. The specification calls for a symmetric algorithm (in which the same key is used for both encryption and decryption), using block encryption of 128 bits and supporting key sizes of 128, 192 and 256 bits. The algorithm was required to offer security of a sufficient level to protect data for the next 20 to 30 years. It had to be easily implemented in hardware and software and had to offer good defenses against various attack techniques. AES has been published as Federal Information Processing Standard (FIPS) 197, which specifies the encryption algorithm that all sensitive, unclassified documents must use.

application

A software component or collection of software components that performs specific user-oriented work (a task) on a computer.

Application Programming Interface

A set of multiple subprograms, data structures and rules for using them that enables application development using a particular language and, often, a particular operating environment. An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.

client/server architecture

An architecture in which the client (usually a personal computer or workstation) is the machine requesting data or services and the server is the machine supplying them. Servers can include microcomputers, minicomputers, or mainframes. The client provides the user interface and may perform application processing. In IBM® Tivoli Monitoring the Tivoli Enterprise Portal is the client to the Tivoli Enterprise Portal Server.

Appendix 4 – SourceText42 (used in the Visual and Blind tasks)

Text_42

Appropriate IBM Tivoli Monitoring RAS1 trace output

IBM Software Support uses the information captured by trace logs to trace a problem to its source or to determine why an error occurred. The reliability, availability, and serviceability (RAS) trace logs are available on the Tivoli® Enterprise Monitoring Server, the Tivoli Enterprise Portal Server, and the monitoring agent. By default, the logs are stored in the installation path for IBM® Tivoli Monitoring.

Sources of other important information

You can collect important information from log files, such as trace or message logs that report system failures. Also, application information provides details on the application that is being monitored, and you can obtain information from messages or information on screen.

Common problem solving

Customers using IBM® Tivoli® Monitoring products or the components of Tivoli Management Services can encounter problems such as missing workspaces or historical data, or a reflex automation script that does not run as expected. In many cases you can recover from these problems by following a few steps. Use the trace settings indicated in these troubleshooting instructions only while trying to diagnose a particular issue. To avoid generating excessive trace data, go back to the default trace settings as soon as the problem is solved.

Diagnosing that workspaces are missing or empty

1. Refresh the Navigator by clicking View > Refresh.
2. Verify that the monitoring agent has been started.
3. In the Tivoli® Enterprise Portal, right-click the Navigator item of the monitoring agent and click Start or Restart
4. Verify that the monitoring agent configuration is correct.
5. If your data is missing in an Oracle Agent workspace, see Resolving Oracle DB Agent problems diagnostic actions.
6. Check that application support has been added.

Glossary

agentless monitoring server

A computer with an OS agent installed that has one or more agentless monitors running on it. Each agentless monitoring server can support up to 10 active instances of the various types of agentless monitors, in any combination. Each instance can communicate with up to 100 remote nodes, which means a single agentless monitoring server can support as many as 1000 monitored systems.

associate

The process of linking a situation with a Navigator item that enables a light to go on and a sound to play for an open event. Predefined situations are associated automatically, as are situations created or edited through the Navigator item pop-up menu. When you open the Situation editor from the toolbar, any situations you create cannot be associated with a Navigator item during this editing session. Close the Situation editor, then open it again from the pop-up menu of the Navigator item with which the situation should be associated.

attribute group

A set of related **attributes** that can be combined in a data **view** or a **situation**. When you launch the view or start the situation, data samples of the selected attributes are retrieved. Each type of monitoring agent has its own set of attribute groups.

display item

An attribute designated to further qualify a situation. With a display item set for a multiple-row attribute group, the situation continues to look at the other rows in the sampling and opens more events if other rows qualify. The value displays in the event workspace and in the message log and situation event console views.

Appendix 5 – TranslationMemory31 (used for SourceText31)

Seg. #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
1	You might not always be able to solve a problem yourself after determining its cause.	You might ~~not always be able~~try to solve a problem yourself after determining its cause.	Es posible que no siempre pueda resolver un problema por sí mismo tras determinar la causa.	~~Es posible que no siempre pueda~~Puede intentar resolver un problema por si mismo tras determinar la causa.	Fuzzy Low
2	For example, a performance problem might be caused by a limitation of your hardware.	For example, a limitation of your hardware might cause a performance problem.	Por ejemplo, un problema de rendimiento podría causarlo un límite de hardware.	Por ejemplo, un límite de hardware podría causar un problema de rendimiento.	Fuzzy Low
3	If you are unable to solve a problem on your own, contact IBM~~®~~ Software Support for a solution.	If you are unable to solve a problem on your own, contact IBM® Software Support for a solution.	Si no puede resolver un problema por sí mismo, póngase en contacto con el soporte de software de IBM para que le proporcionen una solución.	Si no puede resolver un problema por sí mismo, póngase en contacto con el soporte de software de IBM® para que le proporcionen una solución.	Fuzzy High
4	See <a class="xref" href="collectdata_intro_trouble.htm">Logs and data collection for troubleshooting</a> for information on the types of data to collect before contacting Support.	See <a class="xref" href="collectdata_intro_trouble.htm">Logs and data collection for troubleshooting</a> for information on the types of data to collect~~ before contacting Support~~.	Consulte Capítulo 2, “Registros y recopilación de datos para la resolución de problemas”, en la página 3 para obtener información sobre los tipos de datos que hay que recopilar antes de ponerse en contacto con el soporte.	Consulte <a class="xref" href="collectdata_intro_trouble.htm">Registros y recopilación de datos para la resolución de problemas</a> para obtener información sobre los tipos de datos que hay que recopilar~~ antes de ponerse en contacto con el soporte~~.	Fuzzy High
5	Trace data capture transient information about the current operating environment when a component or application fails to operate as designed.	Trace data capture transient information about the current operating environment when a component or application fails to operate as designed.	Los datos de rastreo capturan información transitoria acerca del entorno operativo actual cuando un componente o aplicación no funciona correctamente.	Los datos de rastreo capturan información transitoria acerca del entorno operativo actual cuando un componente o aplicación no funciona correctamente.	Exact
6	IBM Software Support personnel use the captured trace information to determine the source of an error or unexpected condition.	IBM Software Support personnel use the captured trace information to determine the source of an error or unexpected condition.	El personal de soporte de software de IBM utiliza la información de rastreo capturada para determinar el origen de un error o de una condición inesperada.	El personal de soporte de software de IBM utiliza la información de rastreo capturada para determinar el origen de un error o de una condición inesperada.	Exact
7	See <a class="xref" href="tools_trace_trouble.htm">Trace logging</a> for more information about tracing.	See “Trace logging” for more information about tracing.	Consulte “Registro de rastreo” en la página 33 para obtener más información acerca del rastreo.	Consulte “Registro de rastreo” para obtener más información acerca del rastreo.	Fuzzy Low

Seg. #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
8	You can subscribe to e-mail notification about product tips and newly published fixes through the Support portal.	You can subscribe to e-mail notification about product tips and newly published fixes through the Support portal.	Puede suscribirse a notificaciones de correo electrónico sobre consejos del producto y arreglos publicados recientemente a través del portal de Soporte.	Puede suscribirse a la notificación por correo electrónico sobre consejos del producto y arreglos publicados recientemente a través del portal Support.	Machine
9	In the Support portal, you can specify the products for which you want to receive notifications; choose from flashes, downloads, and technotes; and set up to receive email updates.	In the Support portal, you can specify the products for which you want to receive notifications; choose from flashes, downloads, and technotes; and set up to receive email updates.	En el portal de Soporte, puede especificar los productos para los que desea recibir notificaciones; elija entre noticias, descargas y notas técnicas; y configure para recibir actualizaciones de correo electrónico.	En el portal de Soporte, puede especificar los productos para los que desea recibir notificaciones; elija entre noticias, descargas y notas técnicas; y configure para recibir actualizaciones de correo electrónico.	Exact
10	Open the <a class="xref" href="http://ibm.com" target="_blank">http://ibm.com</a> website and select <span class="ph menucascade"><span class="ph uicontrol">Support &amp; downloads</span> &gt; <span class="ph uicontrol">Technical support</span></span>.	Open the <a class="xref" href="http://ibm.com" target="_blank">http://ibm.com</a> website and select <span class="ph menucascade"><span class="ph uicontrol">Support &amp; downloads</span> &gt; <span class="ph uicontrol">Technical support</span></span>.	Abra el sitio web de http://ibm.com y seleccione Soporte y descargas > Soporte técnico.	Abra el sitio web de <a class="xref" href="http://ibm.com" target="_blank">http://ibm.com</a> y seleccione <span class="ph menucascade"><span class="ph uicontrol">Soporte y descargas</span> &gt; <span class="ph uicontrol">Soporte técnico</span></span>.	Exact
11	You can also launch an IBM® support website~~, such as~~ ~~http://www.ibm.com/support/us~~.	You can also launch the IBM® Software Support website~~, such as~~ ~~http://www.ibm.com/support/us~~.	También puede iniciar un sitio web de soporte de IBM, como por ejemplo http://www.ibm.com/support/us.	También puede iniciar el sitio web de Soporte de software de IBM~~, como por ejemplo http://www.ibm.com/support/us~~.	Fuzzy Low
12	In the Quick start page or Support home, click <span class="ph uicontrol">Sign in</span> to sign in or to register if you have not yet registered.	In the Quick start page or Support home, click <span class="ph uicontrol">Sign in</span> to sign in or to register if you have not yet registered.	En la página de inicio rápido o de inicio de Soporte, pulse Regístrese para iniciar la sesión o para registrarse si aún no se ha registrado.	En la página de inicio rápido o de inicio de Soporte, pulse <span class="ph uicontrol">Regístrese</span> para iniciar la sesión o para registrarse si aún no se ha registrado.	Exact
13	In the Notifications area of Support home, click <span class="ph uicontrol">Manage all my subscriptions</span>.	In the Notifications area of Support home, click <span class="ph uicontrol">Manage all my subscriptions</span>.	En el área de Notificaciones del inicio de Soporte, pulse Gestionar todas mis suscripciones.	Notificaciones En el directorio de inicio de área de Soporte, haga clic en <span class="ph uicontrol">Gestionar todas las suscripciones mi</span>.	Machine

Seg. #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
14	In the <span class="ph uicontrol">Subscribe</span> and <span class="ph uicontrol">My defaults</span> tabs, select a product family and continue setting your preferences to specify the information you want in your emails.	In the <span class="ph uicontrol">Subscribe</span> and <span class="ph uicontrol">My defaults</span> tabs, select a product family and continue setting your preferences~~ to specify the information you want in your emails~~.	En los separadores Suscribirse y Mis valores predeterminados, seleccione una familia de productos y continúe configurando las preferencias para especificar la información que desea recibir por correo electrónico.	En los separadores <span class="ph uicontrol">Suscribirse</span> y <span class="ph ucontrol">Mis valores predeterminados</span>, seleccione una familia de productos y continúe configurando las preferencias~~ para especificar la información que desea recibir por correo electrónico~~.	Fuzzy Low
15	If you have not yet added an email address to your profile, click <span class="ph menucascade"><span class="ph uicontrol">My IBM</span> &gt; <span class="ph uicontrol">Profile</span> &gt; <span class="ph uicontrol">Edit</span></span> and add it to your personal information.	If you have not yet added an email address to your profile, click <span class="ph menucascade"><span class="ph uicontrol">My IBM</span> &gt; <span class="ph uicontrol">Profile</span> &gt; <span class="ph uicontrol">Edit</span></span> and add it to your personal information.	Si aún no ha añadido una dirección de correo electrónico en su perfil, pulse Mi IBM > Perfil > Editar y añádala a su información personal.	Si aún no ha añadido una dirección de correo electrónico en su perfil, haga clic en <span class="ph menucascade"><span class="ph uicontrol">Mi IBM</span> &gt; <span class="ph uicontrol">Profile</span> &gt; <span class="ph uicontrol">Editar</span></span> y añadirlo a su información personal.	Machine
16	An encryption algorithm for securing sensitive but unclassified material designed by the National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce.	An encryption algorithm for securing sensitive but unclassified material designed by the National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce.	Un algoritmo de cifrado para proteger el material desclasificado pero sensible diseñado por el NIST (National Institute of Standards and Technology) del Departamento de Comercio de EE.UU.	Un algoritmo de cifrado para proteger el material desclasificado pero sensible diseñado por National Institute of Standards and Technology (NIST) de la U.S. Departamento de Comercio	Machine
17	AES is intended to be a more robust replacement for the <strong class="ph b">Data Encryption Standard</strong>.	AES is intended to be a more robust replacement for the ~~<strong class="ph b">~~Data Encryption Standard~~</strong>~~.	AES está diseñado para ser un sólido sustituto del Estándar de cifrado de datos.	AES está diseñado para ser un sólido sustituto del ~~<strong class="ph b">~~Estándar de cifrado de datos~~</strong>~~.	Fuzzy High
18	The specification calls for a symmetric algorithm (in which the same key is used for both encryption and decryption), using block encryption of 128 bits and supporting key sizes of 128, 192 and 256 bits.	The specification calls for a symmetric algorithm (in which the same key is used for both encryption and decryption), using block encryption of 128 bits and supporting key sizes of 128, 192 and 256 bits.	La especificación llama a un algoritmo simétrico (en el que se utiliza la misma clave para el cifrado y el descifrado), utilizando el cifrado de bloque de 128 bits y que soporta los tamaños de clave de 128, 192 y 256 bits.	La especificación llama a un algoritmo simétrico (en el que se utiliza la misma clave para el cifrado y el descifrado), utilizando el cifrado de bloque de 128 bits y que soporta los tamaños de clave de 128, 192 y 256 bits.	Exact

Seg. #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
19	The algorithm was required to offer security of a sufficient level to protect data for the next 20 to 30 years.	The algorithm was required to offer security of a sufficient level to protect data for the next 20 to 30 years.	El algoritmo debía ofrecer suficiente seguridad para proteger los datos durante los próximos 20 a 30 años.	El algoritmo debía ofrecer suficiente seguridad para proteger los datos durante los próximos 20 a 30 años.	Machine
20	It had to be easily implemented in hardware and software and had to offer good defenses against various attack techniques.	It had to provide an easy implementation in hardware and software and ~~had~~ to offer good defenses against various attack techniques.	Tenía que ser fácilmente implementado en hardware y software y tenía que ofrece una buena defensa contra distintas técnicas de ataque.	Tenía que proporcionar una implementación fácil en hardware y software y ~~tenía que~~ *ofrece* una buena defensa contra distintas técnicas de ataque.	Fuzzy Low
21	AES has been published as Federal Information Processing Standard (FIPS) 197, which specifies the encryption algorithm that all sensitive, unclassified documents must use.	~~AES has been published as~~The Federal Information Processing Standard (FIPS) ~~1~~97~~, which~~ specifies the encryption algorithm that all sensitive, unclassified documents must use.	AES ha sido publicado como Federal Information Processing Standard (FIPS) 197, que especifica el algoritmo de cifrado que deben utilizar todos los documentos sensibles y desclasificados.	~~AES ha sido publicado como~~ Federal Information Processing Standard (FIPS) ~~1~~97~~, que~~ especifica el algoritmo de ~~cifrado~~ encriptación que deben utilizar todos los documentos sensibles y desclasificados.	Fuzzy Low
22	A software component or collection of software components that performs specific user-oriented work (a <strong class="ph b">task</strong>) on a computer.	A software component or set of software components that performs specific user-oriented work (a <strong class="ph b">task</strong>) on a computer.	Componente de software o conjunto de componentes de software que realiza un trabajo específico orientado a usuario (una tarea) en un sistema.	Componente de software o conjunto de componentes de software que realiza un trabajo específico orientado a usuario (una <strong class="ph b">tarea</strong>) en un sistema.	Fuzzy High
23	A set of multiple subprograms, ~~and~~ data structures and ~~the~~ rules for using them that enables application development using a particular language and, often, a particular operating environment.	A set of multiple subprograms and data structures and the rules for using them that enables application development using a particular language and, often, a particular operating environment.	Un conjunto de varios subprogramas y estructuras de datos y las reglas para utilizarlos que habilita el desarrollo de aplicaciones y, a menudo, un entorno operativo en particular.	Un conjunto de varios subprogramas y estructuras de datos y las reglas para utilizarlos que habilita el desarrollo de aplicaciones y, a menudo, un entorno operativo en particular.	Fuzzy High
24	An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.	An API is a functional interface supplied by the operating system or by a separately licensed program that allows an application ~~program~~ written in a high-level language to use specific data or functions of the operating system or the licensed program.	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que un programa de aplicación escrito en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.	Una API es una interfaz funcional suministrada por el sistema operativo o por otro programa bajo licencia que permite que una ~~programa de~~ aplicación escrita en un lenguaje de alto nivel utilice datos o funciones específicos del sistema operativo o del programa bajo licencia.	Fuzzy High

Seg. #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
25	An architecture in which the client (usually a personal computer or workstation) is the machine requesting data or services and the server is the machine supplying them.	An architecture in which the client (usually a personal computer or workstation) is the machine requesting data or services and the server is the machine supplying them.	Una arquitectura en la que el cliente (generalmente un sistema personal o estación de trabajo) es la máquina que solicita los datos o servicios y el servidor es la máquina que los proporciona.	Una arquitectura en la que el cliente (generalmente un sistema personal o estación de trabajo) es la máquina que solicita los datos o servicios y el servidor es la máquina que los proporciona.	Exact
26	Servers can include microcomputers, minicomputers, or mainframes.	Servers can be microcomputers, minicomputers, or mainframes.	Los servidores pueden ser microsistemas, minisistemas o sistemas principales.	Los servidores pueden ser microsistemas, minisistemas o sistemas principales.	Fuzzy High
27	The client provides the user interface and may perform application processing.	The client provides the user interface and may perform application processing.	El cliente proporciona la interfaz de usuario y puede realizar el proceso de las aplicaciones.	El cliente proporciona la interfaz de usuario y puede realizar el proceso de aplicaciones.	Machine
28	In <span class="keyword">IBM® Tivoli Monitoring</span> the <span class="keyword">Tivoli Enterprise Portal</span> is the client to the <span class="keyword">Tivoli Enterprise Portal Server</span>.	In <span class="keyword">IBM® Tivoli Monitoring</span> the <span class="keyword">Tivoli Enterprise Portal</span> is the client to the <span class="keyword">Tivoli Enterprise Portal Server</span>.	En IBM Tivoli Monitoring, Tivoli Enterprise Portal es el cliente del servidor de Tivoli Enterprise Portal.	En <span class="keyword">IBM® Tivoli Monitoring</span> el <span class="keyword">Tivoli Enterprise Portal</span> es el cliente de <span class="keyword">Tivoli Enterprise Portal Server</span>.	Machine

Appendix 6 – TranslationMemory42 (used for SourceText42)

Seg #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
1	IBM Software Support uses the information captured by trace logs to trace a problem to its source or to determine why an error occurred.	IBM Software Support uses the information captured by trace logs to trace a problem to its source~~ or to determine why an error occurred~~.	El soporte de software de IBM utiliza la información capturada por los registros de rastreo para rastrear un problema hasta su origen o para determinar la causa de un error.	El soporte de software de IBM utiliza la información capturada por los registros de rastreo para rastrear un problema hasta su origen~~ o para determinar la causa de un error~~.	Low Fuzzy Match
2	The reliability, availability, and serviceability (RAS) trace logs are available on the <span class="keyword">Tivoli® Enterprise Monitoring Server</span>, the <span class="keyword">Tivoli Enterprise Portal Server</span>, and the monitoring agent.	The reliability, availability, and serviceability (RAS) trace logs are available on the <span class="keyword">Tivoli® Enterprise Monitoring Server</span>, the <span class="keyword">Tivoli Enterprise Portal Server</span>, and the monitoring agent.	La fiabilidad, disponibilidad y servicio (RAS) de los registros de rastreo están disponibles en servidor de Tivoli Enterprise Monitoring, en servidor de Tivoli Enterprise Portal y en el agente de supervisión.	El la fiabilidad, disponibilidad y servicio (RAS) los registros de rastreo están disponibles en la <span class="keyword">Tivoli® Enterprise Monitoring Server</span>, el <span class="keyword">Tivoli Enterprise Portal Server</span>, y el agente de supervisión.	Machine Translation
3	By default, the logs are stored in the installation path for <span class="keyword">IBM® Tivoli Monitoring</span>.	By default, the logs are stored in the installation path for <span class="keyword">IBM® Tivoli Monitoring</span>.	De forma predeterminada, los registros se almacenan en la vía de acceso de instalación para IBM Tivoli Monitoring.	De forma predeterminada, los registros se almacenan en la vía de acceso de instalación para <span class="keyword">IBM® Tivoli Monitoring</span>.	Exact Match
4	You can collect important information from log files, such as trace or message logs that report system failures.	~~You can collect~~Log files can provide important information~~ from~~, such as trace or message logs that report system failures.	Puede recopilar información importante de los archivos de registro, como los registros de rastreo o de mensajes que informan de fallos de sistemas.	~~Puede recopilar~~Los archivos de registro pueden proporcionar información importante~~ de l~~, como los registros de rastreo o de mensajes que informan de fallos de sistemas.	Low Fuzzy Match
5	Also, application information provides details on the application that is being monitored, and you can obtain information from messages or information on screen.	~~Also, application information provides~~ For details on the application that is being monitored, ~~and ~~you can obtain information from messages or information on screen.	Además, la información de la aplicación proporciona detalles sobre la aplicación que se está supervisando, y puede obtener información de los mensajes o información en pantalla.	~~Además, la información de la aplicación proporciona ~~Para ver los detalles sobre la aplicación que se está supervisando, ~~y~~ puede obtener información de los mensajes o información en pantalla.	Low Fuzzy Match

Seg #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
6	Customers using <span class="keyword">IBM® Tivoli® Monitoring</span> products or the components of <span class="keyword">Tivoli Management Services</span> can encounter problems such as missing workspaces or historical data, or a reflex automation script that does not run as expected~~when it should~~.	Customers using <span class="keyword">IBM® Tivoli® Monitoring</span> products or the components of <span class="keyword">Tivoli Management Services</span> can encounter problems such as missing workspaces or historical data, or a reflex automation script that does not run when it should.	Los clientes que utilizan los productos de IBM Tivoli Monitoring o los componentes de Tivoli Management Services pueden encontrar problemas como, por ejemplo, que falten espacios de trabajo o datos históricos, o un script de automatización de reflejo que no se ejecuta cuando debería hacerlo.	Los clientes que utilizan los productos de <span class="keyword">IBM® Tivoli Monitoring</span> o los componentes de <span class="keyword">Tivoli Management Services</span> pueden encontrar problemas como, por ejemplo, que falten espacios de trabajo o datos históricos, o un script de automatización de reflejo que no se ejecuta cuando debería hacerlo.	High Fuzzy Match
7	In many cases you can recover from these problems by following a few steps.	In many cases you can recover from these problems by following a few steps.	En muchos casos, puede recuperarse de estos problemas siguiendo unos pasos.	En muchos casos, puede recuperarse de estos problemas siguiendo tan sólo unos pasos.	Machine Translation
8	Use the trace settings indicated in these troubleshooting instructions only while ~~you are~~ trying to diagnose a particular issue~~specific problem~~.	Use the trace settings indicated in these troubleshooting instructions only while you are trying to diagnose a specific problem.	utilice los valores de rastreo indicados en estas instrucciones de resolución de problemas solamente mientras intente diagnosticar un problema específico.	utilice los valores de rastreo indicados en estas instrucciones de resolución de problemas solamente mientras intente diagnosticar un problema específico.	Low Fuzzy Match
9	To avoid generating excessive trace data, go back to the default trace settings as soon as the problem is solved.	Go back to the default trace settings as soon as the problem is solved, to avoid generating excessive ~~trace~~ data.	Para evitar que se generen demasiados datos de rastreo, vaya a los valores predeterminados de rastreo en cuanto se resuelva el problema.	Vaya a los valores predeterminados de rastreo en cuanto se resuelva el problema, para evitar que se generen demasiados datos ~~de rastreo~~.	Low Fuzzy Match
10	Refresh the Navigator by clicking <span class="ph menucascade"><span class="ph uicontrol">View</span> &gt; <span class="ph uicontrol">Refresh</span></span>.	Refresh ~~the Navigator~~ your browser by clicking <span class="ph menucascade"><span class="ph uicontrol">View</span> &gt; <span class="ph uicontrol">Refresh</span></span>.	1. Renueve Navigator pulsando Ver > Renovar.	Renueve ~~Navigator~~ el navegador pulsando <span class="ph menucascade"><span class="ph uicontrol">Ver</span> &gt; <span class="ph uicontrol">Renovar</span></span>.	High Fuzzy Match
11	Verify that the monitoring agent has been started.	~~Verify that~~Check if the monitoring agent has been started.	2. Compruebe que el agente de supervisión se haya iniciado.	Compruebe ~~que~~si el agente de supervisión se ha~~ya~~ iniciado.	Low Fuzzy Match

Seg #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
12	In the <span class="keyword">Tivoli® Enterprise Portal</span>, right-click the Navigator item of the monitoring agent and click Start or Restart	In the <span class="keyword">Tivoli® Enterprise Portal</span>, ~~right-~~click the Navigator item of the monitoring agent and click Start or Restart	En el Tivoli Enterprise Portal, pulse con el botón derecho del ratón el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar	En el <span class="keyword">Tivoli Enterprise Portal</span>, pulse ~~con el botón derecho del ratón~~ el elemento de Navigator del agente de supervisión y pulse Iniciar o Reiniciar	High Fuzzy Match
13	Verify that the monitoring agent configuration is correct.	Verify that the monitoring agent configuration is correct.	3. Verifique que la configuración del agente de supervisión sea correcta.	Verifique que la configuración del agente de supervisión sea correcta.	Exact Match
14	If your data is missing in an Oracle Agent workspace, see <a class="xref" href="common_cpuoraclediag_trouble.htm">Resolving Oracle DB Agent problems - diagnostic actions</a>.	If your data is missing in an Oracle Agent workspace, see <a class="xref" href="common_cpuoraclediag_trouble.htm">Resolving Oracle DB Agent problems - diagnostic actions</a>.	4. Si faltan datos en un espacio de trabajo del Agente de Oracle, consulte “Resolución de problemas del agente de BD de Oracle - acciones de diagnóstico” en la página 30.	Si faltan datos en un Oracle Agent espacio de trabajo, consulte <a class="xref" href="common_cpuoraclediag_trouble.htm">Resolución de problemas del agente DB de Oracle acciones de diagnóstico</a>.	Machine Translation
15	Check that application support has been added.	Check that application support has been added.	5. Compruebe que se haya añadido el soporte de aplicaciones.	Compruebe que se haya añadido el soporte de aplicaciones.	Machine Translation
16	A computer with an OS agent installed that has one or more agentless monitors running on it.	A computer with an ~~OS~~ agent installed that has one or more agentless monitors running on it.	Un sistema con un agente del sistema operativo instalado en el que se ejecuta uno o varios supervisores sin agentes.	Un sistema con un agente ~~del sistema operativo~~ instalado en el que se ejecuta uno o varios supervisores sin agentes.	High Fuzzy Match
17	Each agentless monitoring server can support up to 10 active instances of the various types of agentless monitors, in any combination.	Each agentless monitoring server can support up to 10 active instances of the various types of agentless monitors, in any combination.	Cada servidor de supervisión sin agentes puede dar soporte a un máximo de 10 instancias activas de los diversos tipos de supervisores sin agentes, en cualquier combinación.	Cada servidor de supervisión sin agentes puede dar soporte a un máximo de 10 instancias activas de los diversos tipos de supervisores sin agentes, en cualquier combinación.	Exact Match
18	Each instance can communicate with up to 100 remote nodes, which means a single agentless monitoring server can support as many as 1000 monitored systems.	Each instance can communicate with up to 100 remote nodes, which means a single agentless monitoring server can support as many as 1000 monitored systems.	Cada instancia se puede comunicar con un máximo de 100 nodos remotos, lo que significa que un solo servidor de supervisión sin agentes puede dar soporte a 1.000 sistemas supervisados.	Cada instancia se puede comunicar con un máximo de 100 nodos remotos, lo que significa que un solo servidor de supervisión sin agentes puede dar soporte a 1.000 sistemas supervisados.	Exact Match

Seg #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
19	The process of linking a situation with a Navigator item that enables a light to go on and a sound to play for an open event.	The process of linking a situation with a Navigator item that enables a light to go on and a sound to play for an open event.	El proceso de enlace de una situación con un elemento de Navigator que permite que se encienda una luz y se reproduzca un sonido para un suceso abierto.	El proceso de enlace de una situación con un elemento de Navigator que permite que se encienda una luz y se reproduzca un sonido para un suceso abierto.	Machine Translation
20	Predefined situations are associated automatically, as are situations created or edited through the Navigator item pop-up menu.	Predefined situations are associated automatically, as are situations created or edited through the Navigator item pop-up menu.	Las situaciones predefinidas se asocian automáticamente, igual que las situaciones creadas o editadas mediante el menú emergente de elementos de Navigator.	Las situaciones predefinidas se asocian automáticamente, al igual que las situaciones creadas o editaras mediante el menú emergente del elemento de Navigator.	Machine Translation
21	When you open the Situation editor from the toolbar, any situations you create cannot be associated with a Navigator item during this editing session.	When you open the Situation editor from the toolbar, any situations you create cannot be associated with a Navigator item during this editing session.	Cuando el usuario abre el editor de situaciones en la barra de herramientas, las situaciones que crea no se pueden asociar con un elemento de Navigator durante esta sesión de edición.	Cuando el usuario abre el editor de situaciones en la barra de herramientas, las situaciones que crea no se pueden asociar con un elemento de Navigator durante esta sesión de edición.	Exact Match
22	~~You must c~~Close the Situation editor, then open it again from the pop-up menu of the Navigator item with which the situation should be associated.	You must close the Situation editor, then open it again from the pop-up menu of the Navigator item with which the situation should be associated.	Debe cerrar el editor de situaciones y volverlo a abrir desde el menú emergente del elemento de Navigator con el que se debe asociar la situación.	Debe cerrar el editor de situaciones y volverlo a abrir desde el menú emergente del elemento ~~de~~ Navigator con el que se debe asociar la situación.	High Fuzzy Match
23	A set of related <strong class="ph b">attributes</strong> that can be combined in a data <strong class="ph b">view</strong> or a <strong class="ph b">situation</strong>.	A set of related <strong class="ph b">attributes</strong> that can be combined in a data <strong class="ph b">view</strong> or a <strong class="ph b">situation</strong>.	Conjunto de atributos relacionados que se pueden combinar en una vista de datos o en una situación.	Conjunto de <strong class="ph b">atributos</strong> relacionados que se pueden combinar en una <strong class="ph b">vista</strong> de datos o en una <strong class="ph b">situación</strong>.	Exact Match
24	When you ~~open~~launch the view or start the situation, data samples of the selected attributes are retrieved.	When you open the view or start the situation, data samples of the selected attributes are retrieved.	Cuando abra la vista o inicie la situación, se recuperan muestras de datos de los atributos seleccionados.	Cuando abra la vista o inicie la situación, se recuperan muestras de datos de los atributos seleccionados.	High Fuzzy Match
25	Each type of <strong class="ph b"></strong> monitoring agent has its own set of attribute groups.	Each type of <strong class="ph b"></strong> monitoring agent has its own set of attribute groups.	Cada tipo de agente de supervisión tiene su propio conjunto de grupos de atributos.	Cada tipo de <strong class="ph b"></strong> agente de supervisión tiene su propio conjunto de grupos de atributos.	Exact Match

Seg #	Source in Text	Source in TM	Translation (original)	Translation in TM	Type of suggestion
26	An attribute designated to further qualify a situation.	An attribute ~~designated~~used to ~~further~~ qualify a situation.	Atributo designado para calificar mejor una situación.	Atributo ~~designado~~utilizado para calificar ~~mejor~~ una situación.	Low Fuzzy Match
27	With a display item set for a multiple-row attribute group, the situation continues to look at the other rows in the sampling and opens more events if other rows qualify.	With a display item set for a multiple-row attribute group, the situation continues to look at the other rows in the sampling and opens more events ~~if other rows qualify~~.	Con un elemento de visualización establecido para un grupo de atributos de varias filas, la situación sigue examinando las demás filas del muestreo y abre más sucesos si otras filas cumplen los requisitos.	Con un elemento de visualización establecido para un grupo de atributos de varias filas, la situación sigue examinando las demás filas del muestreo y abre más sucesos ~~si otras filas cumplen los requisitos~~.	High Fuzzy Match
28	The value displays in the event workspace and in the message log and situation event console views.	The value displays in the event workspace and in the message log and situation event console views.	El valor se muestra en el espacio de trabajo de sucesos y en las vistas del registro de mensajes y de la consola de sucesos de situación.	El valor se muestra en el espacio de trabajo de sucesos y en el registro de mensajes y de consola de sucesos de situación vistas:	Machine Translation

Appendix 7 – Translation instructions for the Visual task

V

Instrucciones

Esta prueba de traducción contiene unas 550 palabras de material IBM (Tivoli).

Para esa prueba, el gestor de proyectos te envía un proyecto (carpeta) de IBM TranslationManager con una **memoria de traducción** que incluye material de una versión anterior del producto y otra memoria con segmentos traducidos por el motor de **traducción automática** de tauyou, entrenado con corpus Tivoli de IBM.

También tienes a tu disposición el texto original en formato impreso para referencia.

Tu tarea consiste en traducir el texto que te han enviado, pensando que después el resultado de tu trabajo será valorado por un revisor profesional.

Appendix 8 – Translation instructions for the Blind task

B

Instrucciones

Esta prueba de traducción contiene unas 550 palabras de material IBM (Tivoli).

Para esta prueba, el gestor de proyectos te envía un proyecto (carpeta) de IBM TranslationManager **sin memoria de traducción**. El texto original en inglés ha sido **pretraducido** al español utilizando una **memoria de traducción** que incluye material de una versión anterior del producto y otra memoria con segmentos traducidos por el motor de **traducción automática** de tauyou, entrenado con corpus Tivoli de IBM.

También tienes a tu disposición el texto original en formato impreso para referencia.

Tu tarea consiste en revisar el texto que te han enviado, pensando que después el resultado de tu trabajo será valorado por un revisor profesional.

Appendix 9 – Research participant release form, Main experiment

RESEARCH PARTICIPANT RELEASE FORM

I voluntarily agree to participate in a series of translation tests for research conducted for the Intercultural Studies Group at the Rovira i Virgili University in Tarragona, Spain.

I understand that this evaluation is being conducted by Carlos da Silva Cardoso Teixeira and will be part of his subsequent Doctoral thesis.

I understand that the evaluation methods that may involve me are:

1. my completion of assessment questionnaires
2. screen recordings of my translation process
3. video and audio-recordings of myself during the translation process
4. eye-tracking (recording of eye movements)
5. my participation in two 20-30 minute post-translation interviews

I grant permission for the interview to be recorded and transcribed, and to be used only by the forementioned researcher for analysis of interview data. I grant permission for the evaluation data generated from the above methods to be published in his thesis and future publications by the Intercultural Studies Group.

I understand that the reports and publications will contain no identifiable information in regard to my name.

____

Signature

____

Full Name

____

Location, Date

Appendix 10 – Research participant release form, Interviews

RESEARCH PARTICIPANT RELEASE FORM

I voluntarily agree to participate in a series of research interviews conducted for the Intercultural Studies Group at the Rovira i Virgili University in Tarragona, Spain.

I understand that this research is being conducted by Carlos da Silva Cardoso Teixeira and will be part of his subsequent Doctoral thesis.

I grant permission for the interviews to be recorded and transcribed, and to be used only by the forementioned researcher for analysis of interview data. I grant permission for the evaluation data generated from the above methods to be published in his thesis and future publications by the Intercultural Studies Group.

I understand that the reports and publications will contain no identifiable information in regard to my name.

____

Signature

____

Full Name

____

Location, Date

Appendix 11 – Metadata in translation tools[21]

The following is a list of possible translation metadata elements that I identified in a previous study when analysing five translation tools (Teixeira 2014b):

- The language pairs involved in the file(s) being translated, usually indicated by country flags or language abbreviations.
- Translation progress statistics, such as the percentage of translated, reviewed or remaining segments.
- The state of segments, including:
    - "translation status" (translated, not translated, automatically propagated, reviewed, pending, approved, etc.);
    - original provenance (whether the translation was typed from scratch or was post-edited from an MT feed or from a TM match: exact, fuzzy match, etc.).
- Terminology suggestions from term bases (glossaries): Typically, text portions identified by the tool as terms are highlighted in the source text, with the corresponding translations and additional information displayed in a separate pane.
- Variables and entities: Similarly to the above, tags, numbers, times, units, etc. are identified and highlighted.
- Type of textual element being translated: These include headings, regular paragraphs, list items, footnotes, table cells, etc. They can typically be indicated through text formatting within the segment, with a letter or code next to the segment, through a preview pane, or a combination of those elements.
- Segment number, line number (in the file), number of characters or words in the source/target segment.
- Typing aids in the form of automatic text (generated either from a predefined list or from glossary or TM matches), which also display as on-screen information.

[21] In this appendix I refer to the metadata that are available within the translation editing environment of translation memory tools, where documents are actually translated. Translation tools usually have separate environments for managing projects, for dealing with files within the projects, for configuring terminology databases, etc. and those environments present their own sets of metadata.

- Automatic indicators for spelling mistakes or other potential editing mistakes (such as tag and number misplacements).
- Indications of whether a segment is the result of two or more segments being manually joined or whether two or more segments were originally a single segment that was manually split.
- Information about translation suggestions,: project-specific, historical (author, date of creation, date of modification, etc.) and linguistic information (fuzzy match levels, differences between source texts, etc.).

This last type of metadata – information about translation suggestions – has been the focus of this thesis and has been called simply "translation metadata" for the sake of simplification.

Translation metadata can be divided into two broad categories: provenance metadata and translation-memory metadata. *Provenance metadata* indicate whether a suggestion comes from a translation memory – and which – or a machine translation engine – and which. *Translation-memory metadata* can be further subdivided into three categories, which I tentatively name as "project-specific", "historical" and "linguistic" metadata. *Project-specific metadata* can include file name, project name, client name and subject domain of the text from which a translation suggestion was produced. *Historical metadata* concern the time and date when a translation segment was created, changed or used; the name of the person who created, modified or used it; and the number of times that segment was used. *Linguistic metadata* indicate the similarities between the text in the source segment being translated and the text in the source segment(s) of the translation memory(ies) from which translation suggestions were produced.