Master

Analysis of the Impact of the Use of Post-editing Guidelines in Raw Machine Translation Output

HIDALGO RIOS, Remedios

Abstract

This master's thesis is related to human post-editing (PE) of raw machine translation (MT) output. It aims to identify the particular differences between post-edited versions with and without guidelines. An experiment with participants was carried out to obtain the relevant information. The test group post-edited the raw MT output with post-editing guidelines, while the control group post-editing without guidelines. A third group carried out a human post-editing evaluation. Different aspects were analysed and studied, for example, the time that the participants needed to perform the post-editing task and its relation with the quality of the output, the number and types of changes made during the post-editing phase in both groups and the post-editing effort. We use the PE guidelines proposed by Marian Flanagan and Tina Paulsen in their study of 2014 "Testing post-editing guidelines : how translation trainees interpret them and how to tailor them for translator training purposes".

Reference

HIDALGO RIOS, Remedios. *Analysis of the Impact of the Use of Post-editing Guidelines in Raw Machine Translation Output*. Maîtrise : Univ. Genève, 2015

UNIVERSITÉ DE GENÈVE
FACULTÉ DE TRADUCTION ET D'INTERPRÉTATION

Remedios Hidalgo Ríos

MA Thesis

Analysis of the Impact of the Use of Post-editing Guidelines in Raw Machine Translation Output

Supervisor: Prof. Pierrette Bouillon
Examiner: Dr. Lucía Morado Vázquez

Department of Translation Technology
Faculty of Translation and Interpreting
University of Geneva
August 2015

I certify that this MA thesis and the research work to which it refers are the product of my own work and that it has not been submitted for a MA degree at any other university. Any ideas or quotations from the work of other people - published or otherwise - are fully acknowledged.

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la *Directive en matière de plagiat des étudiant-e-s,* le *Règlement d'études de la Faculté de traduction et d'interprétation* ainsi que l'*Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction*).

J'atteste que ce travail est le fruit d'in travail personnel et a été rédigé de manière autonome.

Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet.

Je suis consciente que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions.

Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Hidalgo Ríos, Remedios

Genève, 24/08/2015

Acknowledgements

Original text

I would like to thank Pierrette Bouillon for her help and support during the whole adventure at the FTI and especially through this last phase as the supervisor of my MA thesis. I would also like to thank Lucía Morado Vázquez for accepting being the jury of this MA thesis as well for her comments and help. Thanks also to my "guinea pigs" for their selfless participation: without them this work would not exist.

Special and deepest thanks to my family, especially to my parents, for all their love, support, advice and encouragement in all the decisions and stages of my life. Another special mention to my Geneva family: thanks for all of those moments and for all that will come. This adventure would have not been the same without all of you.

Last, but not least, thanks to a very special person for always being there for me. Danke schön.

PE version

Me gustaría dar las gracias a Pierrette Bouillon por su ayuda y apoyo durante toda la aventura en la FTI y especialmente a través de esta última fase ~~es~~ siendo ~~el~~ la supervisora de mi ~~tesis~~ tesina ~~de maestría~~. Me gustaría también dar las gracias a~~l~~ Lucía Morado Vázquez por aceptar ser el jurado de esta ~~tesis~~ tesina ~~de maestría~~, así como ~~para~~ por sus comentarios y ayuda. Gracias también a mis "conejillos de indias" por su participación desinteresada, sin ellos no existiría este trabajo.

El agradecimiento más profundo y especial a ~~Un agradecimiento especial y más profundos a~~ mi familia, en especial a mis padres, por todo su amor, apoyo, consejo y aliento en todas las decisiones y ~~las~~ etapas de mi vida. Otra mención especial es para ~~a~~ mi familia de Ginebra, gracias por todos esos momentos y por todos los que vendrán. Esta aventura no habría sido lo mismo sin todos ~~ustedes~~vosotros.

Por último, pero no menos importante, gracias a una persona muy especial por estar siempre ahí para mí. Danke schön.

Contents

List of Abbreviations

AI: Artificial Intelligence

ALPAC: Automatic Language Processing Advisory Committee

BLEU: Bilingual Evaluation Understudy

CAT: Computer-assisted Translation

CBMT: Corpus-based Machine Translation

CDER: Cover Disjoint Error Rate

CL: Controlled Language

CNGL: Centre of Global Intelligent Content

FAHQMT: Fully-Automatic High-Quality Machine Translation

GALE: Global Autonomous Language Exploitation

MT: Machine Translation

NIST: The National Institute of Standards and Technology

PE: Post-editing

RBMT: Rule-based Machine Translation

SMT: Statistical Machine Translation

TAP: Think Aloud Protocol

TAUS: Translation Automation User Society

TER: Translation Error Rate

WTO: World Trade Organization

List of Figures and Tables

Introduction

The main topic of this master's thesis is the post-editing of machine translation.

Machine Translation is the automation, completely or partially, of the process of translation from one human language into another human language (Arnold et al. 1994). On the other hand, the definition of post-editing is not as fixed as machine translation. One possible definition is the one proposed by O'Brien, who describes post-editing as "the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria" (O'Brien 2011, 197).

This master's thesis is related to human post-editing (PE) of raw machine translation (MT) output, paying special attention to the impact of guideline use during the post-editing phase.

Its main objective is to analyse the impact of post-editing with the use of PE guidelines. Concretely, it aims to identify the particular differences between two post-edited versions. To obtain the information needed, the participants were divided into three groups to carry out different experiments. The first group post-edited the raw MT output with post-editing guidelines (test group). The second group post-edited the raw MT output without any type of guidelines (control group). The third group carried out a human post-editing evaluation. Different aspects were analysed and studied. In particular, the time that the participants needed to perform the post-editing task and its relation with the quality of the output, the number and types of changes made during the post-editing phase in both groups and the post-editing effort. We used the PE guidelines proposed by Marian Flanagan and Tina Paulsen in their study of 2014 "Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes" (Flanagan and Paulsen 2014). In this study, they used the TAUS PE guidelines (please see section 2.7 for further information) before proposing an improved version of them.

The subject of our study arose after reading various works from different authors (e.g. O'Brien; Flanagan and Paulsen 2014; Guerberof 2013; Brunette 2011;

Torrejón and Rico 2012; Plitt and Masselot 2010) discussing post-editing and machine translation. The different authors write about post-editing from different points of view, for example, Guerberof analysed the productivity and the quality in post-editing of outputs from translation memories and machine translation. O'Brien proposed a methodology to measure the correlations between machine translatability and post-editing effort and Rico and Torrejón described the skill-set of the new role of the translator as MT post-editor. In another example, Flanagan and Paulsen evaluated post-editing guidelines to obtain the publishable quality and proposed their own set of guidelines. All these authors arrived at a common conclusion: post-editing should be taken more into account, and translators must adapt to the new age of their field due to the development of new technologies and the aim to substitute human translators, mainly for economic and time related reasons.

The main inspiration for this master's thesis was the previously mentioned article of Flanagan and Paulsen, especially because there is little evidence on post-editing guidelines in the market. The two most popular sets of PE guidelines amongst post-editors in academic and professional markets are the guidelines proposed by the Translation Automation User Society (TAUS), in partnership with the Centre for Global Intelligent Content (CNGL), and the GALE, proposed by the National Institute of Standards and Technology (NIST). This thesis aims to analyse the impact of the use of PE guidelines and identifying the differences when post-editing a raw MT output with and without a set of guidelines.

This thesis is divided into five chapters. Chapter 1 presents a brief history of Machine Translation and an overview of the most popular MT systems. In chapter 2, different aspects of Post-editing are described, e.g. potential definitions, differences between post-editing and translation and revision, evaluation methods and PE guidelines. The methodology (for instance, purposes and objectives, experiment settings and evaluation criteria) used for this master's thesis are explained in chapter 3. The experiment carried out is treated in chapter 4: the design and execution of both pilot test and the experiment as well as its planning and structure are described. Chapter 5 describes the analysis and results.

1. Machine Translation

The aim of this study is to analyse the impact of the use of post-editing guidelines of raw machine translation output. Machine translation is therefore central for this study.

This chapter is divided into two sections. First, in section 1.1, a brief history of machine translation is presented, describing the most important moments since its origin. Second, the different approaches to machine translation currently in use are described in section 1.2.

1.1 Brief History of Machine Translation

Machine Translation can be defined as translation applied through computers: it automatizes, completely or partially, the process of translation from one human language into another human language (Arnold et al. 1994).

The origins of machine translation (MT) can be traced back to the 17^{th} century, when Descartes and Leibniz discussed universal languages and suggested the use of mechanical dictionaries to overcome barriers of language. Nonetheless, the first practical suggestions were not proposed until the 20^{th} century. Patents were proposed independently by Artsouni and Smirnov-Troyanskij in 1933 in France and in Russia respectively. The prototype designed by George Artsouni for a French-Armenian storage device on paper tape was demonstrated in 1937. Smirnov-Troyanskij was convinced that "the process of logical analysis could itself be mechanised" (Hutchins and Somers 1992, 5). He envisaged three stages of mechanical translation: first, a "logical" analysis of words was performed by a source language editor; secondly, the machine transformed the result of the previous analysis (sequences of base forms and syntactic functions) into equivalent sequences in the target language; finally, a target language editor converted the output of the second process, into normal forms in the target language (Hutchins 1994; Hutchins 2006).

However, the actual development of MT can be traced back to 1949, only a few years after the first appearance of computers in 1945 (L'Homme 2008). In particular, Warren Weaver (a director at the Rockefeller Foundation) and Andrew

D. Booth (a British crystallographer) were pioneers in discussing the possibility of using computers for translation. Booth started to explore the mechanization of a bilingual dictionary and began to collaborate with Richard H. Richens (researcher in Cambridge and Director of the Commonwealth Bureau of Plant Breeding and Genetics). Nevertheless, a memorandum from Warren Weaver was the starting point for MT research. Weaver suggested methods for automatic translation using computer techniques, most of which were statistic-related methods (Hutchins and Somers 1992). Due to this idea and its strategic value, numerous research groups all over the world started to invest in this research in the 1950s, including full-time researchers in MT, such as Bar-Hillel at the Massachusetts Institute of Technology (MIT).

In 1954, Leon Dostert, from Georgetown University, in collaboration with the IBM research group, performed the first public demonstration of the feasibility of a MT system with the translation of Russian into English. At the time, these results had little scientific value, but they were sufficient to stimulate funding (at least in the USA) and to inspire new projects in the field (Hutchins and Somers 1992).

A few years later, the positive view of MT was decreasing. Bar-Hillel was the first researcher bringing out doubts about the possibility of having a fully automatic high-quality translation (FAHQT) due to the "semantic barriers" to MT. Instead, he proposed that MT should have less ambitious goals (Hutchins and Somers 1992).

In 1964 funding agencies in the United States formed the Automatic Language Processing Advisory Committee (ALPAC) to carry out an analysis of the quality, costs and prospects of MT. The result of this study was the well-known ALPAC report, published in 1966, in which the realities and limitations of MT were presented. The report was critical, concluding that MT was "slower, less accurate and twice as expensive as human translation" and that "there was no immediate or predictable prospect of useful Machine Translation". It also recommended the development of machine aids for translation and basic linguistic research instead of focusing on machine translation (Hutchins and Somers 1992; Koehn 2010).

Despite the negative results, at the time the report was published, at least three systems were in regular use: at the Wright Patterson USAF base; at OAK Ridge Laboratory of the United States Atomic Energy Commission; and at EURATOM Centre at Ispra, Italy (Arnold et al. 1994).

In the years that followed, research and investment in MT research shrank dramatically and remained as a field of interest for only a handful of groups in Canada (the TAUM group in Montreal, who developed the METEO system), in the USSR (the groups led by Mel'čuk, and Apresian) and in Europe, in particular France and Germany (the GETA group in Grenoble and the SUSY group in Saarbrücken, respectively). At the same time, funding was oriented to research on Computational Linguistics and Artificial Intelligence (AI) (Arnold et al. 1994). It was not until the late 1970s that MT was reconsidered as an important field of research.

The 1980s were marked by the appearance of commercial MT systems in both public and private organizations. Those were years of initiatives as well as rediscovering new paths for old projects. Some good examples of the new wave that MT research experienced are SYSTRAN, descendent of the earliest systems at Georgetown University, and METAL, developed at the University of Texas and later commercialized by Siemens, a private company. Projects such as the EUROTRA, commanded by the European Commission and building on the work of the GETA and SUSY groups in order to create a multilingual MT system, also contributed in the reactivation of the MT research (Hutchins and Somers 1992; Arnold et al. 1994).

In 1992, SDL, a new company focusing on language technologies was founded in the UK by Mark Lancaster. The company has been increasing and developing translation technology since its foundation. In 2002, it acquired the machine translation technology, focusing on Statistical Machine Translation (SMT). Only three years later, in 2005, Trados, the most popular CAT tool among translators was property of SDL. Since its origins, SDL has led the innovation in language technology and has become one of the biggest and most popular companies in the field in a little over 20 years (SDL 2014).

Little by little the public became more aware of the importance of MT as well as the companies and research groups that continued developing commercial and open source systems. Machine translation is still a very productive field, reflected in numerous research groups (e.g. Machine Translation group in the School of Computing at Dublin City University and the Statistical Machine Translation Group at the University of Edinburgh) trying to make the most of their resources to continue the evolution of MT.

In the next section, machine translation evolution will be described. It will start with the linguistic systems, the origin of MT systems (section 1.2.1) and will continue with the most popular approach nowadays, the statistical model (section 1.2.2.1)

1.2 Machine Translation Systems

MT systems can be classified in different ways. When MT systems started to be developed, they could follow a traditional classification depending on the languages involved in the translation. We can distinguish between bilingual systems, working only with a pair of languages, and multilingual systems, developed to work with more than two languages (Hutchins and Somers 1992). This classification is no longer used because nowadays systems are designed to be multilingual. Also, systems can be classified as uni-directional, translating only in one direction, or bi-directional, translating in both senses.

Nowadays, MT systems are classified depending on their architecture: rule-based (RBMT) and corpus-based (CBMT), depending on whether they work with linguistic knowledge (Arnold et al. 1994). RBMT systems can be categorised as direct or indirect.

Figure 1 represents the most popular and common types of MT systems that will be described in the next section: rule-based systems (direct and indirect) and the statistical approach.

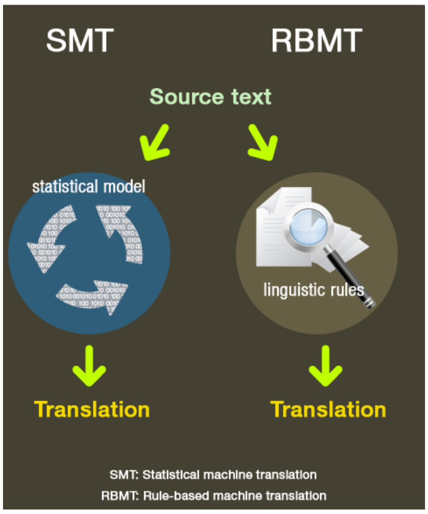


Figure 1: MT systems described in this chapter

1.2.1 Rule-based Systems

The Vauquois triangle (see Figure 2) explains the three main rule-based approaches (direct, transfer and interlingua) and their complexity.

The model is represented by a pyramid, where the vertical direction illustrates the amount of effort necessary for analysis (left side) and generation (right side). In the horizontal direction, the amount of effort needed for transfer is represented (Trujillo 1999). The least complex systems are located at the bottom of the pyramid and the most complex on the top, that is direct systems are located at the bottom, transfer approach in the middle and interlingua at the top. According to this model, more complex systems require higher effort in analysis of source and target language; however, conversion between languages requires lower effort.

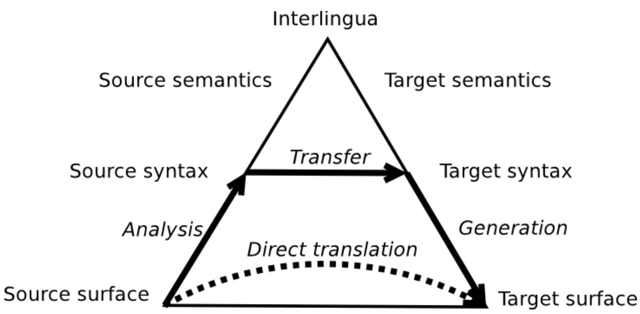


Figure 2: Vauquois triangle

These systems will be described from the least complex to the most. Direct systems will be described in first place, followed by the indirect approaches: transfer and interlingua.

1.2.1.1 Direct Systems

Direct systems are the first generation of MT systems, and were developed in the 1950s. They are very simple and without any deep grammatical knowledge: there is an input in a source language A which is directly processed into a target language B, and there is no intermediate phase in the translation process. Due to their architecture, direct systems are typically designed to work in only one direction and with one pair of languages (Jurafsky 2000).

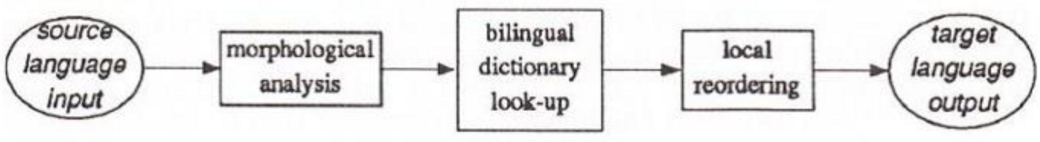


Figure 3: Direct MT system (Hutchins and Somers 1992, 72)

The translation process (see Figure 3), starts with a lexical analysis to carry out a morphological analysis and disambiguation of the source language words. In the next step, these source language words are translated into their word equivalences in the target language using a bilingual dictionary. This step is followed by the application of local reordering rules to obtain a better quality target language output (Arnold et al. 1994).

While this approach may seem simple, direct systems have some positive aspects. They are robust and will give a translation in every case, which will be useful to get the gist of the text; they can work highly effectively with specialized dictionaries and simplified texts that can be translated word-to-word. It is important to mention that the quality of the output will be better when the languages involved are linguistically similar, e.g. the combination Spanish-Italian will work better than Spanish-Japanese (Arnold et al. 1994; Bouillon 2012).

There are also some disadvantages. The comprehension is minimal and there is only a lexical analysis carried out. Due to this fact these systems will have typical lexical mistakes, result of a local lexical analysis and a word-to-word translation (Jurafsky 2000; Bouillon 2012).

There are still various direct systems available on the market, for example Reverso localize. This platform was created in 2012 and nowadays it is one of the most popular systems to localise websites due to its accessibility to all types of users. It is not only used to localise websites, the platform also offers the possibility of translating different types of materials, i.e. documents, websites, apps and data, very easily and in a large number of different languages (Reverso-Softissimo, 2013).

1.2.1.2 Indirect Systems

A different type of rule-based system is the indirect system (second generation). These systems are more sophisticated than the direct approach and are linguistically motivated.

The main difference with direct systems is that the indirect approach includes an intermediate phase in the translation process (hence "indirect", in contrast to "direct"). They go further in the analysis to extract a complete representation of the source text that will help to generate the text in the target language (Hutchins and Somers 1992). This new phase requires that the MT system contains linguistic knowledge in both source and target languages, encoded in a grammar (Arnold et al. 1994).

Two types of indirect system will be explained in the following sections: transfer and interlingual systems.

a) Transfer Systems

There are three independent steps in the process of a transfer system: analysis, transfer and synthesis (see Figure 4).

During the first step, the parser analyses the input with the source language grammar. This approach involves a more detailed analysis. Transfer systems do not carry out only a lexical analysis, as in direct systems; they also perform a syntactic analysis, which disambiguates not only words but also grammatical relations.

The transfer phase, also called the intermediate phase, is a language-dependent step. During this process, the source language and target language representations are matched using transfer rules (Bouillon 2012). The system maps the words and the syntactic structures of the pair of languages involved in the process. The output in the target language is the result of the synthesis step using a target language grammar (Arnold et al. 1994; L'Homme 2008).

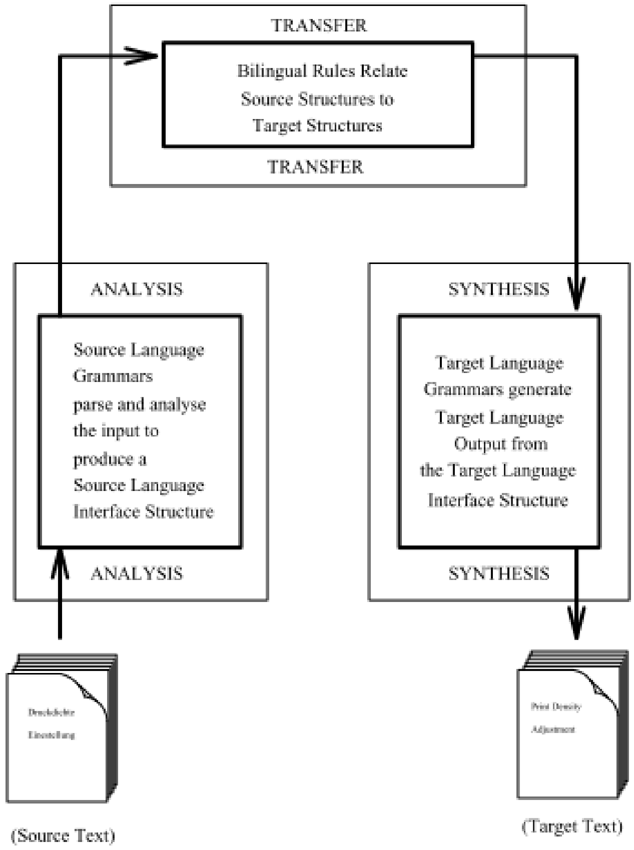


Figure 4: Transfer system (Arnold et al. 1994, 68)

An important strength of transfer systems is that the analysis phase conducted is more detailed that the one performed in direct systems; as a result, there are fewer disambiguation mistakes. In addition, it is possible to process all syntactic differences. The result could be a Fully-Automatic High-Quality Machine Translation (FAHQMT) output if the domain is limited enough to write translation rules that capture all possibilities, for example in the translation of weather forecasts or medical systems. Nowadays, these systems are the only possible means to obtain a FAHQMT output for domains that require high translation accuracy (Arnold et al. 1994; Koehn 2010).

Nevertheless, while very useful with a concrete topic and a specific pair of languages, these systems also have some disadvantages. To introduce new languages to the system, it is necessary to introduce the grammar and create transfer rules for each new pair of languages. This task is complicated and time consuming, requiring specialist in every pair of languages added (L'Homme 2008; Bouillon 2012).

Systems such as SYSTRAN started using the direct approach and have ended up as transfer systems. Another example of these systems is Lucy Software (Comprendium), based on METAL system (Fernández Guerra 2000; Bouillon 2012).

```
SLCAN  «grand»
TLCAN  «high»
PRF    41022
TESTS  ((XFR-AST-TEST :MOD-CAN «temps»))

SLCAN  «grand»
TLCAN  «tall»
PRF    10022
TESTS  ((XFR-AST-TEST :MOD-TYN HUM))

SLCAN  «grand»
TLCAN  «big»
TAG    GV
PRF    10000
```

Figure 5: Example of transfer rules for Lucy (Bouillon 2012 "Cours 5: Le système Lucy (Comprendium)", 15)

b) Interlingual Systems

The second type of indirect system is the interlingual approach, where the main steps are the analysis and the synthesis. The particularity of these systems is that there is no transfer phase but an interlingual representation (see Figure 6).

The system analyses the source text with different processes, using linguistic and pragmatic knowledge to create an intermediate representation that contains all the information required to generate the target text. This representation reflects at the same time the source and target texts and is language-independent (as opposed to transfer systems). Theoretically, due to the interlingual representation, adding new pairs of languages to the system could require less effort than in the transfer systems (Hutchins and Somers 1992; Arnold et al. 1994).

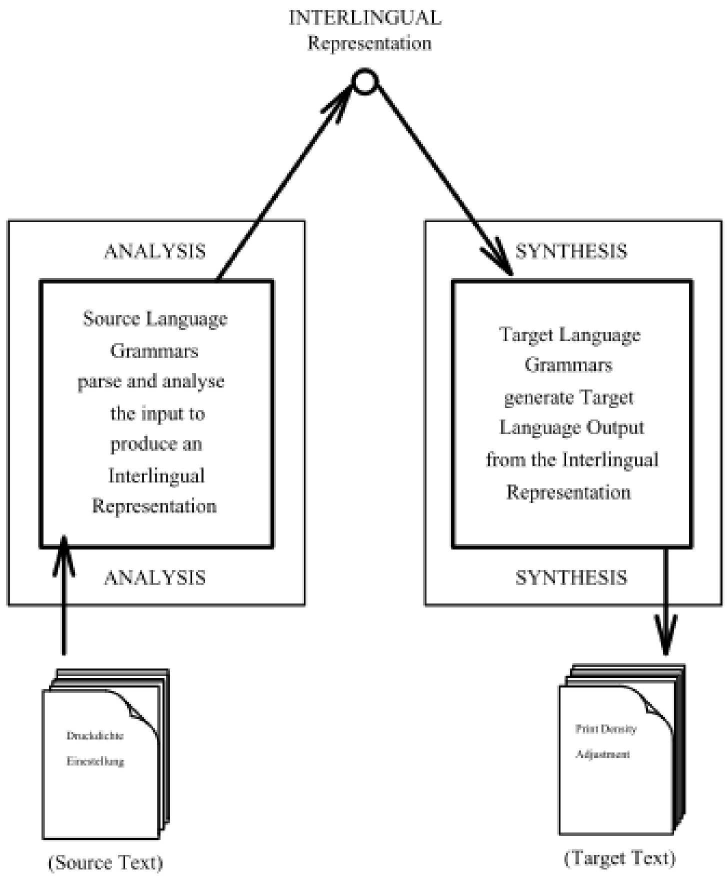


Figure 6: Interlingual system (Arnold et al. 1994, 79)

The main theoretical advantage of the interlingual approach is that a new language can be added easily to the system, as mentioned before, so it could work with more than just a pair of languages without a lot of effort. Actually "in the past, the intention or hope was to develop an interlingual representation which was truly 'universal' and could thus be intermediary between any natural languages" (Hutchins and Somers 1992, 73). Nowadays researchers are less ambitious and restricted to limited domains and languages. This means that in practice the incorporation of a new pair of languages will require creating the grammars of the new languages and defining their characteristics, processes that will require significant effort in every case, especially in the case of languages that are linguistically similar. For instance, the process will be more complex when adding Chinese to a Spanish-French system that adding Italian; regardless, the problem of defining the language will remain (Hutchins and Somers 1992; Arnold et al. 1994).

Despite the complexity of overlapping concepts from different languages to build the interlingua, this approach is more attractive for multilingual systems because no transfer rules are required. However, even if they are still appealing from a developing point of view, research groups are no longer that interested in these systems; instead, statistical machine translation systems are nowadays the main focus of research.

1.2.2 Corpus-based Systems

Corpus-based systems appeared during the 1950s and 1960s with Warren Weaver and the Georgetown system. During the 1990s, with the increasing importance of the internet and the massive amount of data that can be found online, the potential of this approach increased massively. Nowadays, the two main corpus-based systems are: example-based and statistical systems. However, in this work only the statistical system will be described because it is the type of system used during the evaluation of this master's thesis.

1.2.2.1 Statistical Systems

The first prototypes or ideas about statistical systems are not as new as might be expected: it was one of the first techniques used during the 1950s but abandoned during the 1960s to focus on linguistic systems days (Hutchins 2006). They have gained popularity among all types of users, thanks, mainly, to online translation systems developed by Google and Microsoft, the more extended statistical systems to have been developed in recent years, and other MT systems that can be found online.

The appearance of open-source systems also contributed to the development of statistical machine translation. Moses is the most representative example of open-source implementation of the statistical approach. Moses technology is used in teaching and research and, increasingly, in commercial systems. One of the most recent applications of this system is its integration in two post-editing platforms, the results of two European projects: Casmacat and MateCat (Moses 2013). MateCat is the system used for the evaluation presented in this thesis, and will be described in section 3.2.1.

Statistical systems rely on probabilistic and statistical models; this implies that the systems need to be trained on large quantities of parallel data (from which the systems learn how to translate the segments) and monolingual data (from which the systems learn what the target language should look like) (Trujillo 1999; Moses 2013).

The aim of the most common statistical systems is to translate a source sentence by searching in a corpus for the target sentence that has the highest probability of being the translation of the source sentence. The results taken from the corpus are based on probabilities that are based on two different models: a language model containing monolingual information and a translation model containing bilingual information (Trujillo 1999; Koehn 2007).

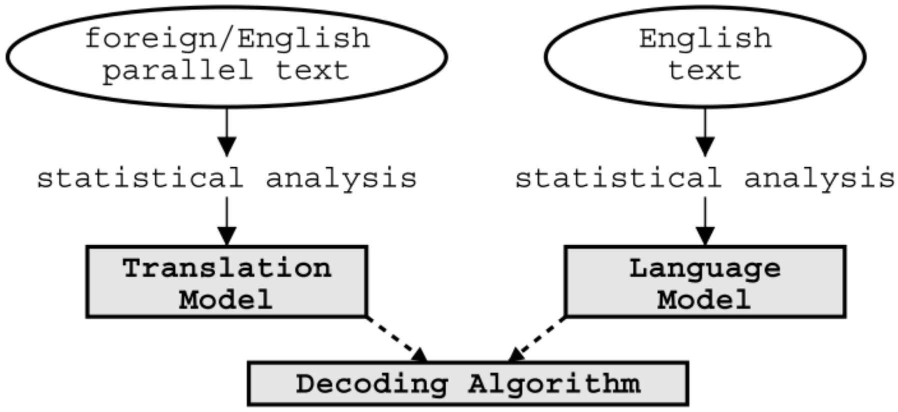


Figure 7: Representation of Language and Translation models (Koehn 2007)

In statistical machine translation, a series of models replaces the set of rules employed by a machine translation system to transform the source sentence into a target sentence (Koehn 2010).

A language model measures the probability of the appearance of words and strings of words (n-grams) of a sentence in a monolingual target-language corpus. The aim of this module is to analyse the linguistic accuracy of the output, choose the best possible linguistic combination, and to optimize fluency in the target text.

A translation model measures how faithful the translation is. The objective is to determine which translation, for a word or string of words of a given sentence in a source language, will be the most probable by searching through an aligned bilingual corpus. Two other points are used to determine the probability of this result: fertility (how many words could be used as a translation) and distortion (the probability of the position of a target word within a sentence).

This type of system has several advantages: it is easy to implement, it needs a specialist in computer science rather than a linguist; there are various open source tools, so it can be cheap to create; as there is no linguistic knowledge involved, linguistic resources, like dictionaries; or specialists, like translators, are not needed.

However, there are not only positive points for these systems but also disadvantages: the quality of the translation depends on the quality and size of the corpora, as with the example-based systems; they work better with close languages and the results may still not be hugely accurate. Non-systematic errors can appear, which are not always easy to post-edit; it is difficult to improve or change the results because not always the user can add lists of vocabulary or try to use pre-editing rules; further, significant informatics resources are needed to implement any change (Bouillon 2012).

1.3 Conclusion

The first part of this chapter has given a brief review of the history of machine translation, giving special attention to the most important milestones of machine translation development and commenting the ups and downs during the decades.

In the second part of the chapter we can find an overview of the main and most popular machine translation approaches that can be found. In the next chapter an overview of post-editing will be presented.

2. Post-editing

The goal of this master's thesis is to analyse the influence of post-editing (PE) guidelines on the process of post-editing raw MT output. This chapter describes the following aspects of post-editing. It presents possible definitions of PE (section 2.1), strategies (2.2), differences between translation and revision (sections 2.3 and 2.4 respectively), how to measure the PE effort (section 2.5) and list the skills that a post-editor should have (section 2.6). Sections 2.7 and 2.8 address PE guidelines and different evaluation methods respectively.

2.1 What is post-editing? Some possible definitions

The term "post-editor" was introduced in the 1950s by Erwin Reifler, who gave the following description to the task of a post-editor: "to select the correct translation from the possibilities found by the computer dictionary and to rearrange the word order to suit the target language" (Krings and Koby 2001).

Post-editing is a recent field, and it is closely related to the applications of machine translation. It is usually understood as the changes made by a human to a machine translated output in order to make the text acceptable for its intended purpose (Krings and Koby 2001). However, due to its relative newness, there are diverse definitions for the concept of post-editing.

Wagner (1985) defines post-editing as a correction of a text that has been already pre-translated rather than a correction "from scratch". Allen (2003) defined the task as the edition, modification and/or correction of a text that has been pre-translated by a MT system. O'Brien (2011) went further than Allen and described post-editing as "the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria".

According to this last definition, guidelines are required to perform this task. We would like to focus on this definition carrying out an experiment with candidates and analysing the impact of the use of guidelines while post-editing a raw MT output.

2.2 Post-editing Strategies

Wagner (1985) proposed two post-editing strategies: full post-editing and rapid post-editing. The aim of full post-editing is to obtain a text with a quality comparable to a human-translated text. Rapid post-editing focusses on correcting the most important errors to make the text comprehensible and accurate, without placing a lot of importance on the style.

The choice of one strategy or the other will depend on the time available to post-edit the text and the quality of text required, e.g. full post-editing will need more time but will provide a better quality.

Krings and Koby (2001) and Allen (2003) followed the classification proposed by Wagner. However, in the 1990s, Allen incorporated to the classification the term “minimal post-editing”, which is also referred to as “post-editing at a minimum” (Allen 2003). The main problem with minimal post-editing is to know which errors should be corrected; this choice is going to vary from one person to another, no matter if it is a translator, a post-editor or a reviser. One possible solution is to work with some specific guidelines to establish the limits of this subjectivity.

Robert (2010) described two different strategies of post-editing: *post-édition brute* and *post-édition évoluée*. The main difference between them is that the first one is the revision, modification, correction, adaptation, reorganization and rereading of a pre-translated text by a MT system, while in the second the raw text is “*produit par un processus qui associe diverses technologies de TA et de TAO*” (Robert 2010, 3), that is, the text is going to be processed by various computer-assisted tools, in particular CAT tools (translation memories, terminological databases) and rule-based MT systems.

In the next sections post-editing will be differentiated from other practices, such as translation and human revision, with which post-editing shares some characteristics.

2.3 Post-editing vs. Translation

In the previous sections, some definitions of post-editing as well as different strategies to post-edit have been described.

Post-editing and translation are two tasks that should not be confused, even if it can seem that there is some appearance of similarity during their process, reality shows that they are different.

O'Brien (2002) reflected the differences between translation and post-editing on the practical level, instead of on cognitive processes. Post-editors work with three different texts (source text, pre-translation and post-edited version); while translators only have two (source text and translation). In addition, she talks about the quality that they have to achieve: translators have to create a text with a high quality but post-editors have to adapt the quality of the target text depending on the needs of the customer and/or the purpose of the text.

As mentioned above, according to Wagner (1985) the post-editor has to face a pre-translated text instead of a blank page, which is what the translator has to deal with. This implies that translators have the possibility of being more creative and free while selecting terms and words, while post-editors are restricted by the raw machine translation text that they have in front of them, which may be of variable quality.

From a different point of view, Krings and Koby (2001) demonstrated with their study that cognitive processes occurring during the tasks of post-editing and translation are different. The source text comprehension is different in each case.

2.4 Post-editing vs. Revision

The key contrast between post-editing and revision is that post-editing is the correction of a text translated by a MT system and revision is the correction of a human translation (Wagner 1985).

Wagner (1987, 73) described the main differences between post-editing and revision:

a) the raw translation is peppered with silly mistakes which no human translator would ever make, and which have to be corrected;
b) you cannot trust the terminology; even if it is right, it has to be checked, and if it is wrong, the right term has to be found;
c) machine translation systems cannot cope with quotations or references to other documents; here again, the post-editor has to do as much research as for a human translation.

Following the idea that revision is the correction of a human translation and post-editing is the correction of a MT output, Krings and Koby (2001) differentiated post-editing from revision by proposing that the types of error that the reviewer faces differ from those ones that the post-editor will find. Human translations are the product of a human brain and this means that a person can make an error once, but, after a feedback, probably this particular error is not going to be repeated. On the other hand, raw machine translation outputs are produced by a MT system, which needs rules and/or corpus, dictionaries, etc. to create the translation of a source text. The differences lie in the frequency, repetitiveness and types of errors. For example, a person can mistranslate a word once, but, when this word has been corrected once after having received a feedback, probably this mistake is not going to be repeated; however, the system is going to repeat the same error all the times this word appears. This means that the person that is going to post-edit a text should be aware that the quality is not going to be the same as the one that a human translation should have (Krings and Koby 2001).

Vasconcellos (1987b, 411) explains this difference as follows:

"With post-editing, the focus is on adjusting the machine output so that it reflects as accurately as possible the meaning of the original text. In other words, the emphasis is on an ongoing exercise of adjusting relatively predictable difficulties, rather than on the discovery of any inadvertent are minor and local, are more frequent than in traditional revision".

2.5 Post-editing Effort

In order to measure post-editing effort, Krings and Koby (2001) introduced three different levels of effort that should be taken into account: temporal, cognitive and technical post-editing efforts.

Temporal effort is the most easily measured of the three levels. It is also the most important one in terms of cost, and thus the one that will be prioritised commercially. However, the time required to post-edit a text will depend on the type of errors: the time needed to correct a punctuation error will differ from that required to address an ungrammatical sentence construction. This point is strongly linked with the cognitive effort, the second type of post-editing effort described by Krings and Koby. It cannot be observed directly, but can be measured depending on the errors that have to be corrected: the most complex errors require greater cognitive activity and vice versa. In order to observe cognitive effort, Krings and Koby used the "Think-Aloud Protocol" (TAP)[1], where the post-editor has to verbalise what is going through his or her mind while working. O'Brien also suggested alternative methodologies for measuring post-editing effort as the use of a tool, such as *Translog*, to record keyboard monitoring (O'Brien 2004). During one of her studies, O'Brien used a special computer monitor to collect data from the eye-tracking[2]. This technique consists on tracking the eye movements, number of fixations and length of time spent in fixations to measure while the post-editor is working (O'Brien 2011). "Fixations" are defined as: "eye movements which stabilize the retina over a stationary object of interest" (Duchowski 2003, 43). According to this definition, the more difficult the text, the longer will be the fixations.

Temporal and cognitive PE efforts depend directly on the quality of the text that is post-edited: both efforts will be lower for high quality machine translation and higher for low quality machine translation (Krings and Koby 2001).

The third effort proposed by Krings and Koby is the technical post-editing effort. This type of effort is based on the technical operations (e.g. deletion,

---

[1] For further information about this method, please refer to Krings and Koby 2001.

[2] For further information on eye tracking-related research in translation, please refer to Mees et al. 2009; Göpferich et al. 2008.

insertion, reordering) needed to implement changes in order to improve the quality of the machine translated text. Technical effort should be differentiated from cognitive effort because, even if the technical operations are controlled by cognitive processes, an error can be easily identified and a correction strategy easily developed, but its technical implementation can be more demanding to perform.

O'Brien (2004) explained that post-editing effort can be reduced adopting some strategies. For instance, she proposed the use of machine translation combined with translation memories to reduce the post-edit workload. Second, machine translation quality might be improved by integrating updated specific MT dictionaries or more adequate corpora, depending on whether it is a rule-based or statistical machine translation system. Third, the source text may be written following controlled language (CL) rules to improve its *translatability* (Mitamura et al. 1998; Nyberg et al. 2003; EAMT/CLAW 2003; Aikawa et al. 2007).

Vasconcellos and Léon (1985) supposed that an experienced post-editor should be able to post-edit between 4 000 and 10 000 words per day or 4 pages, with rapid post-editing, according to Wagner (1985).

2.6 Skill-set for Post-editing

Post-editing is a recent domain, which is why there are still gaps regarding post-editors' skills. To address these gaps, some authors discuss the integration of post-editing modules in the programs of Translation' studies. Post-editing is still in the centre of discussion in both professional and academic contexts.

O' Brien (2002) and Robert (2013) discussed the importance of specialized training for post-editors. Robert is in charge of a course to make university students aware of the importance of post-editing. O'Brien proposed a course content including an introduction to post-editing, machine translation and controlled languages; in addition she recommends a post-editing module with specialised translation, terminology, IT and basic programming. These skills are the ones thought necessary for a post-editor.

According to different authors (Krings and Koby 2001; Rico and Torrejón 2012), the best candidates to work as post-editors are translators. McElhaney and Vasconellos (1988, 143) said "in addition to the professional expertise that a translator brings to the task, it is essential that he has good keyboard skills, quick reactions to the challenges that arise in the text, and a flair for solving problems creatively".

Wagner (1987) suggested that the skills of a translator differ in some ways to those of a post-editor. She cited the following requirements for post-editors:

- Excellent knowledge of the source language
- Perfect command of the target language (=mother tongue)
- Specialized subject knowledge
- Word processing expertise
- Tolerance

She agreed that the three first requirements are not exclusive to post-editors; in fact they are the same for translators. Word processing can belong, at least partially, to translators and will involve (according to Vasconcellos) good keyboard skills, efficient cursor control and the ability to program macros. However, the last requirement, the tolerance for machine translation errors, is specific to post-editors.

Rico and Torrejón (2012) agreed that translators are best suited to the role of post-editor, even if there are some differences between their duties. In their proposal they shared some requirements of the skill-set proposed by Wagner (1987) and also of the modules that should become part of the post-editor training proposed by O'Brien (2002). In addition, they mentioned (see Figure 8) the "core competences" in order to be able to achieve the expectations of the customer, to deal with subjectivity issues while post-editing and to accept the quality of the text, even if this is low.

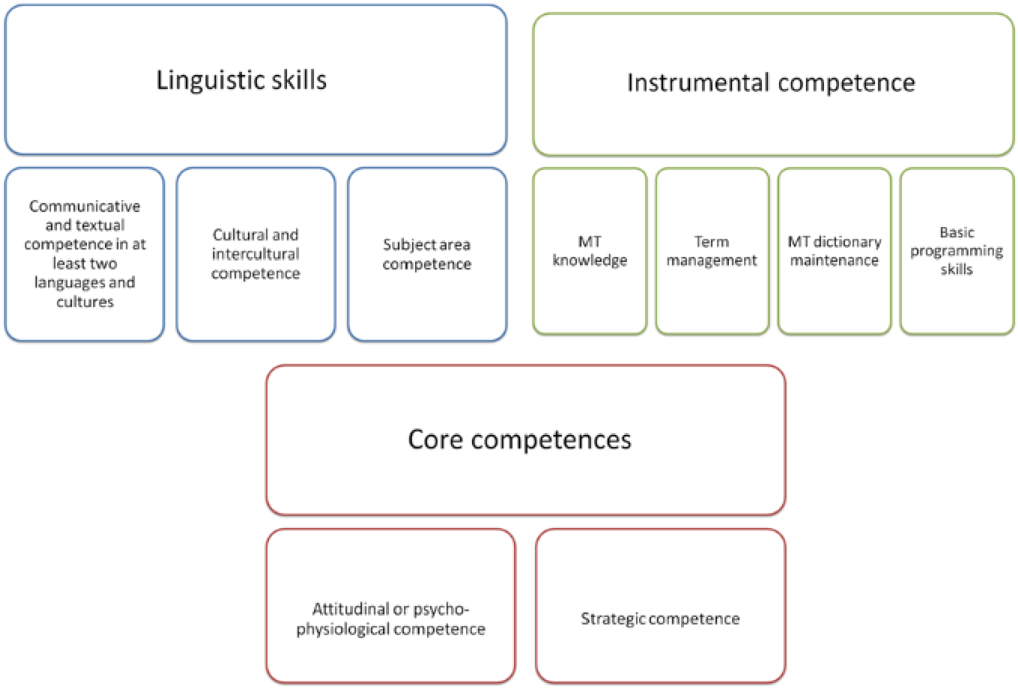


Figure 8: PE skills and competences (Rico and Torrejón 2012, 169)

Krings and Koby (2001), also proposed that translators are the best ones to become post-editors, but for the reason that the primary linguistic skill needed in both roles is revision. This skill is gained by experience, so the experienced revisers could also become post-editors.

Robert (2013) distinguished three types of competences that a post-editor needs. First, she mentioned the mental agility that a post-editor should have to be able to perform his/her role. It is worth mentioning that she understands post-editing as work involving different CAT tools:

- Reviewing the sentences proposed by the translation memory, previously translated by translators (100% accuracy).
- Updating the sentences partially translated by the translation memory (from 75% to 99%).
- Post-editing the sentences produced by a MT system (from 0% to 74%).
- Reading the final product to harmonize and finalise it.

Related to the first competence and the need to work with CAT tools, Robert (2013) defined the second ability of a post-editor as being able to adapt to new technologies and to work with them. The third skill that she considered indispensable is project management: post-editors need to be organised and efficient. Some of the project management tasks associated to post-editing are computer related (e.g. update software and terminology databases), administrative (e.g. bills, prepare cost estimates) and interpersonal (e.g. follow-ups of the project, keep the customer updated).

2.7 Post-editing Guidelines

In the section 2.1, different definitions of PE from different authors were proposed, nevertheless, if we go back to the definition suggested by O'Brien (2011): "the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria", it can be seen that she mentioned the importance of the use of guidelines in order to perform the post-editing task.

Guzmán (2007) does not propose a set of guidelines but he thinks that a certain level of "artificial" language should be tolerated if the text is comprehensible and grammatically correct. If the text follows these two main goals, the post-editing process should be reduced to correcting only errors related to grammar and syntax, misspellings and punctuation and mistranslations. Reducing the types of corrections will save time, effort and money.

Wagner (1985, annexe 1) gives some rules to perform the rapid post-editing in the Commission in Luxembourg. They are the following: read the original text in first place; always make reference to the original text; only make a correction if the sentence is wrong or nonsensical, if there is enough time, also correct the sentences that are ambiguous; resist the temptation to delete everything and rewrite; do not hesitate too long over any particular problem; do not worry if the style is not elegant; do not post-edit text with unfamiliar topics; use only rapid research aids; do not worry about rapid post-editing low quality. They seem to be quite strict except one that recommends saving time being a bit "creative" or to use a colloquial expression, even if normally this case would not be acceptable, just because the aim is to transmit the information.

The National Institute of Standards and Technology (NIST), in the context of the Global Autonomous Language Exploitation (GALE) program, and the Translation Automation User Society (TAUS), in partnership with the Centre for Global Intelligent Content (CNGL), are the authors of the two sets of PE guidelines most used among post-editors: GALE and TAUS post-editing guidelines respectively.

The GALE PE guidelines (Thicke 2013) aim to edit as little as possible in order to have an output with a correct meaning and using an understandable language:

- Make the MT output have the same meaning as the reference human translation - no more, no less.
- Make the MT output as understandable as the reference.
- Capture the meaning in as few edits as possible using understandable English. If words/phrases/punctuation in the MT output are completely acceptable, use them (unmodified) rather than substituting something new and different.
- Punctuation must be understandable, and sentence-like units must have sentence ending punctuation and proper capitalization. Do not insert, delete, or change punctuation merely to follow traditional optional rules about what is “proper.”

The TAUS[3] platform offers different recommendations to reduce the effort and level of post-editing and also suggests two different types of PE guidelines depending if the expected quality is a “good enough” quality or a quality similar to human translation.

A “good enough” quality will be understandable and accurate even if it the style and the syntax are not the correct ones:

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.

[3] TAUS recommendations and guidelines can be consulted here: https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

- Use as much of the raw MT output as possible.
- Basic rules regarding spelling apply.
- No need to implement corrections that are of a stylistic nature only.
- No need to restructure sentences solely to improve the natural flow of the text.

If the desirable quality should be equal to human translation (i.e. the text is comprehensible, accurate and has a correct syntax, grammar and style) the guidelines that should be followed are these:

- Aim for grammatically, syntactically and semantically correct translation.
- Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms".
- Ensure that no information has been accidentally added or omitted.
- Edit any offensive, inappropriate or culturally unacceptable content.
- Use as much of the raw MT output as possible.
- Basic rules regarding spelling, punctuation and hyphenation apply.
- Ensure that formatting is correct.

These last sets of guidelines have been used by different authors like O'Brien and de Almeida (2010), who used the GALE PE guidelines in their study while Flanagan and Paulsen (2014) made an experiment following the TAUS guidelines and they even proposed a new set of guidelines to achieve a quality similar to the human translation. As our experiment will be based on the study carried out by Flanagan and Paulsen, we will talk more in detail about the guidelines proposed in this study in section 3.2.5.

2.8 Post-editing Evaluation

There are different ways to measure the quality of a text that has been translated with a MT system and post-edited by a human: automatic and manual evaluation.

2.8.1 Automatic Post-editing Evaluation

There are some metrics to measure the differences between the MT output and a reference translation in terms of editing steps.

One of the possible metrics is the Edit-Distance, or Levenshtein Distance, a string similarity measure defined as follows: “the minimum number of editing steps — insertions, deletions, and substitutions — needed to match two sequences” (Koehn 2010, 224). The distance between two strings in terms of Levenshtein is the number of edit operations needed to transform one word to another. Permitted edit operations are insert, delete and replace. Each operation comes with a cost, e.g. the cost of adding one letter or deleting one can be 1 and the cost of replacing a letter is either 1 or 0 depending on whether the letters are equal or not. Generally, Edit-Distance breaks down the problem to many sub-problems and solves them successively. Here the sub-problems are the edit distances between prefixes of words (Levenshtein 1965).

For example, the edit distance between "dog" and "do" is 1 (deletion) and the edit distance between cat and act is 2 (replacements). Costs can be modelled in many ways. One could, for instance, assign lower replacement costs to letters that are close by on a keyboard, for example “m” is more likely to be mis-typed as “n” than as “q”. Therefore, replacing “m” by “n” is a smaller edit distance than by “q” (Levenshtein 1965).

The main disadvantage of the Levenshtein distance is that it operates usually on character level. If we were to compute the edit distance of two bi-grams whereby in one of them the word order is changed, the Levenshtein distance gives the number of character operations required to transform one string to the other, i.e. it will provide the cost of deleting and re-inserting the misplaced word. This movement is usually the cut-and-paste function performed by the human post-editor while working on the text (Levenshtein 1965; Koehn 2010).

Other examples of automatic metrics are the translation error rate (TER) and the cover disjoint error rate (CDER). They are based on the Levenshtein distance but add the jump or block movement as a possible editing step. TER tries to find the shortest sequence of editing steps, a computationally demanding task. CDER is simpler, due to its relaxation of alignment requirements.

One possible disadvantage of using these metrics is that the error measured will not represent the real performance of the system evaluated. This is the case when the system output may be an acceptable translation, but the translation differs from the reference translation (Koehn 2010).

A different and more sophisticated approach based on basic counting is the human translation error rate (HTER). This metric is inspired by the human evaluation (see section 2.8.2). This method is described by Koehn (2010, 238) as follows:

"HTER is defined as the number of editing steps divided by the number of words in the acceptable translation."

A disadvantage of this method may be that the HTER score can increase because the number of editing steps may involve the attempts while the post-editor is trying to find the best solution.

2.8.2 Manual Post-editing Evaluation

As opposed to automatic post-editing evaluation, manual evaluation is performed by a human.

According to Hutchins and Somers (1992) there can be two different methods for linguistic evaluation. The first one will be to evaluate the quality of the output. The tests of the quality proposed by Hutchins and Somers (1992, 63) are:

- Fidelity or accuracy: the extent to which the translated text contains the "same" information as the original
- Intelligibility or clarity: the ease with which a reader can understand the translation
- Style: the extent to which the translation uses the language appropriate to its content and intention.

The second method is to count manually each addition, deletion, substitution and modification and calculate the percentage of errors in the whole text.

The main disadvantage of this approach is the time required to perform the task. In addition, the second method cannot be completely objective, mainly because it is carried out by humans and the assessment of what is or is not a mistake will vary from one person to another (Hutchins and Somers 1992; Koehn 2010).

2.9 Conclusion

In this chapter we have done a general tour through the field of post-editing. Different definitions were described to give an idea of what is the purpose of the task, as well as the different levels that can be performed. The differences with other close subjects, such as translation and revision, and skill-set needed to be a post-editor were described. To end the chapter, we focused on a more technical part and we explained different post-editing guidelines that are available nowadays and different methods to evaluate post-editing.

In the next chapter the methodology used for the experiment carried out will be explained.

3. Methodology

This chapter describes the methodology used for this research, namely the analysis of the impact of using post-editing guidelines. It will first explain the purpose and objectives of the evaluation (section 3.1) and it will continue with the evaluation model and settings selected for this study (section 3.2). Finally, the evaluation criteria will be described in section 3.3.

3.1 Purpose and Objectives

Post-editing is a recent and very popular topic at present, with numerous studies and articles on this subject. The main inspiration for this study is the work of Flanagan and Paulsen (2014). In this study, the authors carried out an experiment with university students in order to be able to test the TAUS PE guidelines (see section 2.7) while post-editing two different types of texts. After analysing the results, Flanagan and Paulsen proposed a new set of guidelines for post-editing for publishable quality (see section 3.2.5).

The aim of this thesis is to analyse the differences between post-editing a raw machine translation output while using or not using post-editing guidelines. In order to achieve this goal, the study is divided into two parts. During the first part of the experiment, two groups of candidates will carry out post-editing tasks. The machine translation system used is MateCat (see section 3.2.1). In addition, the opinion of the candidates about post-editing and their satisfaction while working in a PE task will be collected and analysed. Once this will be finished, the differences in terms of quality and time needed to post-edit with the PE guidelines or not will be analysed. The different types of errors and changes made will be also analysed during the second part of the study.

3.2 Experiment Model and Settings

In the next sections, the settings for this study will be described. Section 3.2.1 will describe the post-editing platform chosen to perform the experiment. The subsequent sections, 3.2.2, 3.2.3 and 3.2.4 will treat respectively the platform and language settings and the requirements for the text and participants. Section 3.2.5 will describe the set of guidelines chosen for this experiment. In section 3.2.6, additional settings, for example which browsers were used, will be described.

3.2.1 Post-editing Platform

The post-editing platform selected was MateCat, a statistical open-source web-based and CAT tool developed by a consortium composed of the international research center FBK (Fondazione Bruno Kessler), Translated srl, the Université du Maine and the University of Edinburgh (MateCat 2014).

MateCat works with an advanced Moses[4] integration and the biggest collaborative translation memories. Moses (see section 1.2.2.1) is an open-source statistical machine translation project that started in 2005 as a successor of another project called Pharaoh[5].

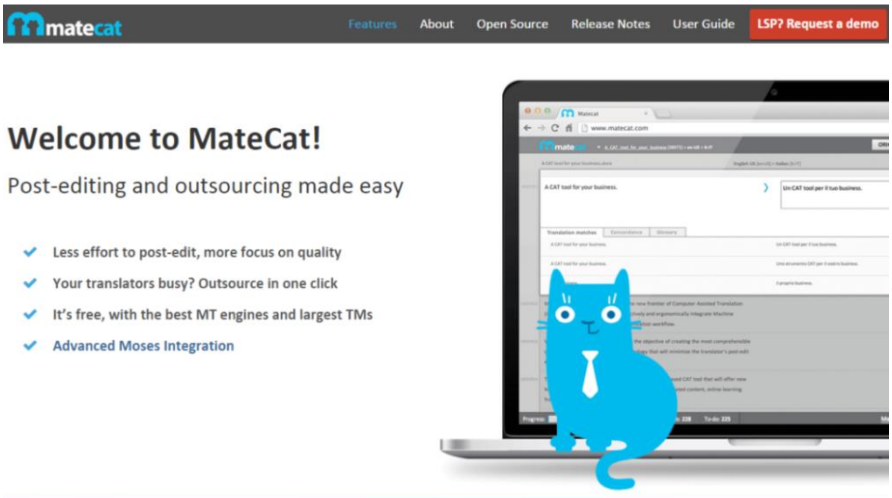


Figure 9: MateCat feature page
http://www.matecat.com/features/ [Accessed 18/06/15]

There are different tabs on the feature page of MateCat (see Figure 9). General information about the platform and also the members of the research group can be found under *About*. Under the *Open Source* tab, there is information about the project (publications, reports, etc.). MateCat offers also the possibility of

[4] For further information, please see http://www.statmt.org/moses/index.php?n=Main.HomePage [Accessed 18/06/15]
[5] For further information about the project, please see http://www.isi.edu/licensed-sw/pharaoh/ [Accessed 16/06/15]

downloading the MateCat package, even though it is a web-based system, in case the user would like to work offline and/or customize the system. An installation guide can also be found on this page. All the updates and news can be found under *Release Notes*.

There is very useful information under the *User Guide* tab, as shown in Figure 10. This page contains all the documentation. Different user guides are available, as well as an introductory video that will help the user to get familiarised with the platform.

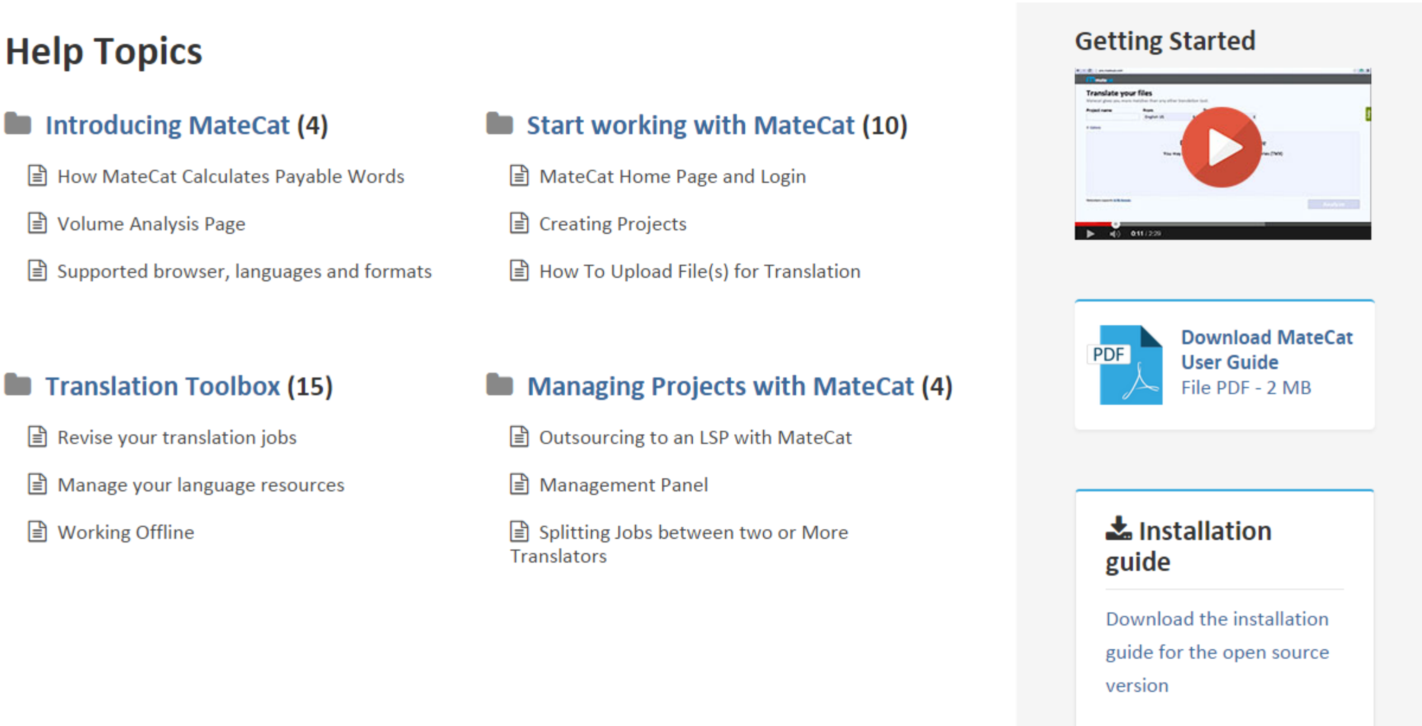


Figure 10: User Guide tab
http://www.matecat.com/support/ [Accessed 18/06/15]

MateCat offers two possibilities to translate. The user can translate the documents or outsource the projects to a translation company. In the case of this master's thesis, the translations were not outsourced; the participants translated the text. Post-editing is also available. The interface is easy to work with and very intuitive. At the moment of writing this dissertation (June/July 2015) it is only available in English.

There is no need to create a special account to work with MateCat; only a Gmail account is needed to log in. The user can log in at any stage of the process, even though it is recommended to log in while creating the project.

From the homepage (see Figure 11), it is possible to create the project and define its parameters, the working languages and the subject. Regarding the subject, the platform offers a list of different domains in order to have a more accurate output. The user can also select the memory that should be used, the user can also upload it in TMX format or create a personal translation memory for a specific project. The platform has been updated recently and now it is possible to add a particular MT engine, the ones used by default are the systems from Google Translate and Microsoft Translator.

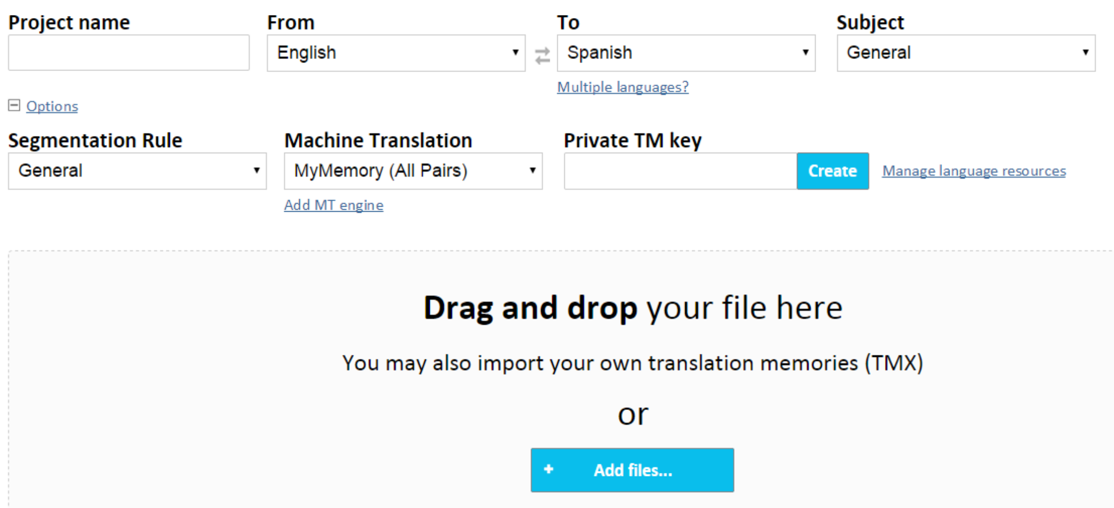


Figure 11: MateCat homepage

https://www.matecat.com/ [Accessed 18/06/15]

Once all the settings are chosen, the files that need to be translated can be uploaded. The system supports 58 different file formats. After carrying out an analysis, the user is ready to translate.

During process of translation and post-editing the different segments are opened automatically and the user can make all the changes required. When the user has finished post-editing the segment, he/she has to confirm the translation of the segment and the next segment will be opened. The user can go back to the translated segments to make more changes if needed.

The original text and the translation can be downloaded at any time in the translation window. It is also possible to download the translation in formats compatible with SDL Trados and OmegaT.

Amongst the features offered by MateCat is the option of evaluating the translation. Once the translation has been completed and all the segments have been confirmed, the user can go to the revision step. In this part of the process, the user can select the types of error, summarised in Figure 12, which can be found in each segment.

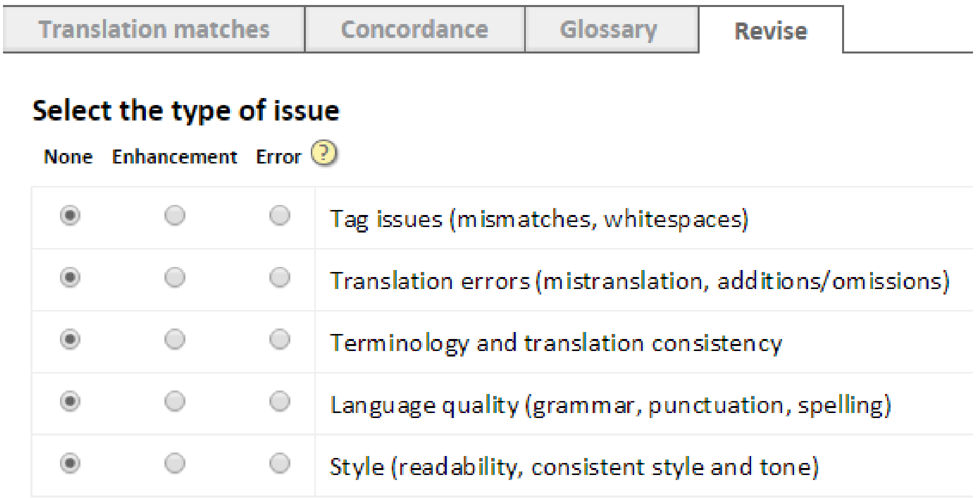


Figure 12: Types of errors proposed by MateCat

There is also an Editing log (see Figure 13) that contains information about each segment, for example, the time the translator worked on it, the changes made compared with the translation proposal by the system, or an estimation of the PE effort.

Secs/Word	Job ID	Segment ID	Words	Suggestion source	Match percentage	Time-to-edit	PE Effort
16.2	97401	50944515	4.00	Machine Translation	85%	01m:04s	57%

Segment	WORLD TRADE REPORT 2014
Suggestion	MUNDIAL DEL COMERCIO INFORME 2014
Translation	INFORME DEL COMERCIO MUNDIAL 2014
Diff View	~~MUNDIAL~~ INFORME DEL COMERCIO ~~INFORME~~ MUNDIAL 2014

Figure 13: Example of editing log

The Editing log gives important and useful information to analyse the post-editing phase (MateCat 2014, Support-Editing log). This information is summarised in Table 1.

Secs/Word	Seconds spent per word in the segment
Job ID	MateCat job ID
Segment ID	ID of the segment. You can click on it to open the segment in the Translation Editor page
Words	Number of words in the segment
Suggestion source	Source of the first suggestion
Match percentage	Match percentage of the first suggestion. It is always 85% by default for Machine Translation suggestions
Time-to-edit	Time necessary to edit the segment, measured by adding the time used to perform multiple edits on the same segment
PE Effort	Post-editing effort as a percentage. This indicates the efforts in editing machine translation suggestions. It is measured by the word edit distance between the first match provided and the final translation
Segment	Original source segment
Suggestion	First MateCat suggestion
Translation	Final translation approved by the translator
Diff View	Differences between first suggestion and final approved translation highlighted with track changes

Table 1: Information given in the Editing log of MateCat

In general, it can be said that MateCat offers some extra functions, e.g. post-editing interface, which made this tool more complete in comparison with other web-based statistical machine translation systems.

3.2.2 Platform and Language Settings

The experiment will be executed with the online version of MateCat. The basic features (e.g. editing log, revision phase) available in this version are enough to carry out the assignment.

The subject chosen for this task is *General*. The platform has a pull-down menu with multiple options but our text could fit partially in different categories (see section 3.2.3). We tested different texts and the results were the same, so it was decided to choose this subject in order to make the task easier.

In this specific task, the users did not use a specific translation memory or a different machine translation engine from the default version.

The working languages chosen for this evaluation are English and Spanish, being English the source language and Spanish the target language. There were only two possible source languages choices: French or English, because those are the languages offered at the University of Geneva as possible working languages with Spanish. In addition, it was a requirement to find an original text, instead of a translation, and it is easier to find original texts in English. Another good reason was that as this master's thesis is written in English and all the interfaces of the program are only available in English.

As mentioned above, Spanish was chosen as the target language. The main reason was because a translator usually works towards his/her native language and, in this case, Spanish is the mother tongue of the researcher. Another reason is that it was easier to find candidates with this particular pair of languages.

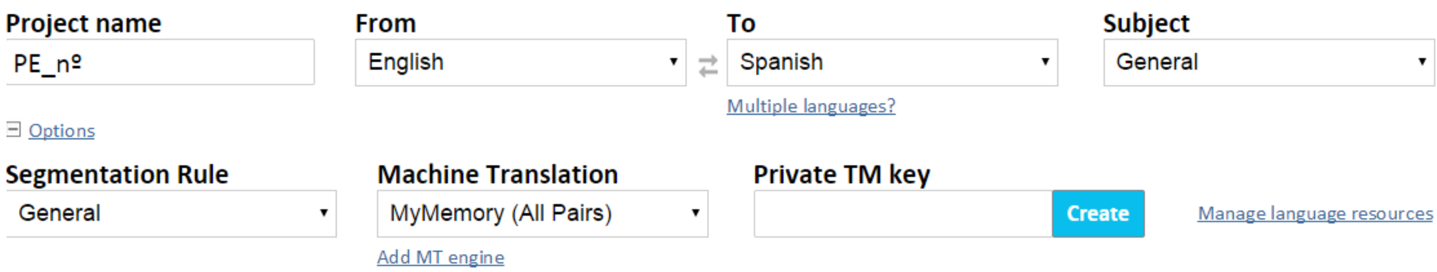
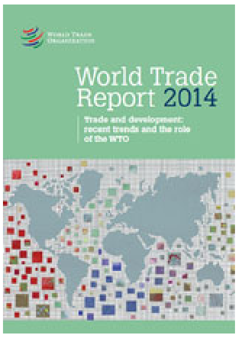


Figure 14: Settings for the evaluation

3.2.3 Text

The text selected for this study was an extract from the World Trade Report of 2014, the annual report of the Word Trade Organization (WTO). The text is available online at the official website of the Organization, https://www.wto.org/, under the sections *Publications* and *Economic research*, in the menu *Documents, data and resources* (see Figure 15).



RESEARCH AND ANALYSIS

World Trade Report 2014

Trade and development: recent trends and the role of the WTO

The World Trade Report 2014 looks at how four recent major economic trends have changed how developing countries can use trade to facilitate their development. These trends are the economic rise of developing economies, the growing integration of global production through supply chains, the higher prices for agricultural goods and natural resources, and the increasing interdependence of the world economy. The Report also looks into what role the WTO plays.

Figure 15: World Trade Report 2014 appearance in the WTO website

The reasons for this choice were related to the language, the type of text and the subject.

The text is originally in English, which means that even if there are versions in French and Spanish available, the original was written in English. The candidates will not be working with a translated text but with the original. Besides, the fact that the official WTO translation of the text in Spanish is also available online left open the possibility of comparing the quality with an example of the real publishable quality expected by the Organization. As mentioned before in the previous section, it was a requirement of the researcher to keep some language uniformity in this master's thesis, an additional reason to choose English against French.

It was decided to select a semi-informative text in order to obtain a raw MT output of an acceptable quality according to BLEU score. The aim was to challenge the translation system with some specialised terminology that was not into MateCat database or, at least, that did not have a high number of appearances on its database so the system could have problems in translating the terminology. On the other hand, we did not want to choose highly specialised text that might make the task unnecessarily difficult for the participants.

BLEU score was calculated to have an objective score of the quality of the raw MT output. BLEU is one of the most popular metrics to evaluate the quality of a machine translation output. The score is calculated considering matches of n-grams between the MT output and a reference translation and goes between 0 and 1. Scores closer to 1 represent a higher quality (Koehn 2010). Please refer to Koehn 2010, pp. 226-228 for further information.

Regarding the subject, as mentioned before, a specific domain was needed. It was decided to choose a domain with which the participants were familiar. For this reason it was decided to select either an economic or legal text, the main topics of translation lectures offered at the University of Geneva. Another requirement was to make sure that all the participants involved in the study had undertaken sufficient courses in this field to be familiar with the type of text and not just the domain. Due to these requirements it was decided to choose economy as the topic of the text.

The characteristics of the chosen text are the following:

- Domain: economy
- Source: extract of the annual report of the WTO
- Number of words: 1 032
- Number of segments: 50
- BLEU Score[6]: 0.77

3.2.4 Participants

When planning the different experiments that were to be carried out for this master's thesis, it was considered to create a test and a control group in order to perform two different types of post-editing tasks, with and without PE guidelines. Nevertheless, when recruiting the participants for this task, two different types of profiles were available. This fact was the starting point of the

[6] See section 5.6

creation of a new group to conduct the second part of the study, the human post-editing evaluation of the PE versions against the raw MT output (see section 4.4).

The final number of participants was eleven (see Table 2). For the first part of the task, two groups of four people were composed. The second part was carried out by two candidates. In addition, an extra participant was needed to perform a pilot test before starting with the official study (see section 4.3).

Task		Number of participants
Pilot test	Pilot test participant	1
Post-editing	Test group (PE with guidelines)	4
	Control group (PE without guidelines)	4
PE evaluation	Evaluators	2

Table 2: Number and organisation of the participants

The groups that carried out the first part of the project, the post-editing tasks, were formed by students or recent graduates of the Faculty of Translation and Interpreting (FTI) at the University of Geneva. All of the participants were master's students in their second year. They were writing their master's thesis or were recently graduates during the task execution. They all belonged to Translation Technologies and/or Specialised Translation master's specialisations; it was not a prerequisite to look for students from a specific master's specialisation. The only requirement regarding the master's specialisation was that they must have taken at least one of the Translation Technologies lectures. The main reason for this requirement was to have a list of candidates familiarised with the study program of the FTI and also to be sure that they had some knowledge regarding translation technologies. Regarding their professional careers, most of them had some experience as freelance translators but not as post-editors.

The second part of the project was carried out by participants who did not belong to the FTI or the University of Geneva. This group was composed of participants with a degree in Translation and Interpreting and with a master related to Translation in both cases. They had more experience as freelance translators than the other candidates but were not as familiar with translation technologies as the participants chosen for the post-editing tasks. A special and mandatory characteristic was that all participants must have Spanish as a native language.

The information provided by the participants in the different questionnaires will be analysed in chapter 5.

3.2.5 PE Guidelines

The post-editing guidelines that were used are a set of guidelines proposed by Flanagan and Paulsen in their study of 2014 "Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes" (Flanagan and Paulsen 2014). In this study they used the post-editing guidelines proposed by TAUS (see section 2.7) for a publishable quality with a group of three students in order to analyse how the participants understood and interpreted the different guidelines.

In our case we used the PE guidelines proposed by the authors of the article mentioned above. These guidelines were conceived to provide publishable quality translation, as opposed to "good enough" quality, and are the following:

1) Use as much of the raw MT output as possible, *but*:
2) Ensure the target reader perfectly understands the content of the target text
3) Ensure the target text communicates the same meaning and message as the source text
4) Ensure target-text language is appropriate, but do not restructure sentences solely to improve the flow of the target text
5) Aim for a grammatically, syntactically and semantically correct target text

6) Ensure key terminology is correctly translated (considering context and the client's list of preferred terminology, if available)

7) Ensure any untranslated terms belong to the client's list of "Do Not Translate" terms, if available

8) Edit any offensive, inappropriate or culturally unacceptable content for the target reader

9) Apply basic rules regarding spelling and punctuation

10) Ensure the same ST tags are present and in the correct positions in the target text

3.2.6 Additional Settings

In addition to the parameters described in the previous sections, some additional decisions were taken.

When planning the experiment, it was considered to carry out the evaluation at the Faculty of Translation and Interpreting, concretely in the computer lab located in the sixth floor. All the participants would have worked at the same time in the same space and the researcher would be present to answer any possible doubt or problem that may arise in order to give more consistency to the evaluation. The use of a program like BB FlashBack[7] to record information like mouse clicks and keystrokes to record the time that the participants took to perform the task was also considered. However, when recruiting the participants, most of them had problems of availability, a fact that made it impossible to carry out the evaluation as initially planned.

Due to this logistic problem, it was decided that the participants would carry out the task remotely; they also had the option of doing it with the researcher if they preferred. They had MateCat instead of another MT platform because, as described in section 3.2.1, it is web-based and it was still possible to have a record of the time required to perform the task and the PE effort.

[7] http://www.bbsoftware.co.uk/BBFlashBack/Home.aspx [Accessed 18/06/15]

Google Chrome and Safari were the browsers selected because, besides the fact that they are two of the most popular internet browsers, they are the only browsers on which MateCat works.

As an extra step, Gmail accounts were created for every participant in order to be able to create the project with MateCat. The different email addresses followed the structure below, where "X" was the corresponding user number for each participant:

- Gmail account: user.X.PE@gmail.com
- Password: UsuarioXPE

3.3 Evaluation Criteria

Different variables were evaluated in this master's thesis. The data required for the analysis were collected with two different methods: the task that the candidates carried out and the questionnaires that they completed once finished the practical assignment.

The information gathered from the questionnaires helped to analyse various characteristics (see section 4.2.3 for further information about the questionnaires). The main topic to consider was the vision of the candidates regarding post-editing, they gave some information about their experience, and how did they feel about this practice and also if they thought that post-editing can be harmful for translators. Related to this point, they were asked if they thought that post-editing guidelines were useful or not and if they could change the way of post-editing a text. In addition, test group participants were asked if they were able to follow and understand the instructions of the task and the guidelines while the control group participants were asked if they thought that having some post-editing guidelines could have been useful. The information given in the questionnaires about the types of errors that they considered more difficult to post-edit was compared with the types of errors marked in MateCat by the participants.

The practical assignment was divided into two different tasks. The aim of the first experiment, the post-editing tasks, was to collect different kind of data.

The editing log provided by MateCat, described in section 3.2.1, provided information about the time that the participants needed to perform the task, more specifically, it was possible to have the total amount of time needed and the time that the participant used to edit each segment. Similarly, the average post-editing effort needed for each segment and the overall average post-editing effort was available.

During the second part of the task, the post-editing evaluation, the type and minimum number of modifications that were made in order to convert the MT output into a translation with publishable quality, was measured. MateCat also calculated an automatic HTER that was combined with the manual task to obtain the final result. In the manual task, the evaluators counted the insertions, deletions, substitutions and shifts manually to have an exact number of changes. In addition, the different types of errors were identified. We compared the subjective results from the questionnaires and the revision phase of MateCat. During the revision phase in MateCat, the participants selected the type of errors that they found in each segment. Additionally, the accuracy was measured using a five degree scale.

The variables that were evaluated, the type of evaluation (objective and/or subjective) and the method used to analyse the different results are summarised in table 3.

VARIABLES EVALUATED	OBJECTIVE EVALUATION	SUBJECTIVE EVALUATION
Opinion about PE	-	Specific questionnaire
Opinion about PE guidelines	-	Specific questionnaire
PE Accuracy	-	Human PE Evaluation
PE effort	MateCat editing log	Specific questionnaire
Quality of the MT output	BLEU score	Specific questionnaire
Time	MateCat editing log	Specific questionnaire
Evaluation of the PE outputs	MateCat HTER	Human PE evaluation
Types of errors in the MT output	MateCat Revision	Specific questionnaire

Table 3: Variables evaluated during the experiment

3.4 Conclusion

During this chapter the methodology has been described. Different aspects have been taken into account, in particular purposes and objectives of this study, task and platform parameters as well as the evaluation criteria and PE guidelines chosen.

The website of MateCat has been presented. The different parameters that can be chosen in order to create a translation project have been described, as well as the main features and strengths.

In the next chapter the evaluation design and its execution will be explained. The creation of the materials and the execution of the pilot test will be also described.

4. Experiment Design and Execution

This chapter will explain the design and execution of the experiment that was carried out during this master's thesis. The chapter will start with a description of the different tasks (section 4.1). It will continue with the documentation in section 4.2. Hereafter, in section 4.3, the pilot test and its execution will be explained. To conclude the chapter, a description of the execution of the evaluation that the participants have to complete will be treated in section 4.4.

4.1 Experiment Structure

The experiment planned for this master's thesis was structured into two main tasks in order to obtain two post-edited versions of raw MT output and to have a human post-editing evaluation of these post-edited outputs. These two tasks have been designed for two types of participants: post-editors and evaluators.

The experiment was divided into two different post-editing assignments to be carried out by the test and control groups. Both groups post-edited the same text previously translated by MateCat. The final outputs were exported and inserted manually by the researcher in different Excel spreadsheets. There was one Excel spreadsheet for each post-editor (see section 4.2.4).

The post-editing evaluation was done by two evaluators, who evaluated the results of the post-editing experiment. The Excel spreadsheets from the post-editors groups were divided between the evaluators. One of the evaluators completed the Excel spreadsheets corresponding to the test group and the second evaluator, the ones from the control group. The original idea was that the evaluators completed the Excel spreadsheets of both groups to analyse the differences between the results of the two evaluators. However, due to availability reasons, this last step did not take place. Additionally, they performed the revision step on MateCat (see section 3.2.1) to select the different types of issues present in the different segments of the text.

Post-editors from both groups and evaluators completed two questionnaires: a set of background questions and another questionnaire more specific to their particular task.

In addition, documentation for each task of the evaluation was created in order to support the participants while going through their assignments. The different materials will be explained in the next sections.

4.2 Documentation

As mentioned above, various types of materials were prepared: information for participants, instructions and questionnaires. In addition, an Excel spreadsheet was developed for the human evaluation. All documentation was revised and approved by the responsible member of the Faculty before the evaluation took place.

It was decided to write all the materials in Spanish, due to all the participants have Spanish as native language.

4.2.1 Information for Participants

All the terms and conditions of the participation in this experiment were documented in the document called *Información para los participantes* (see Annexe II). In this document, general information was provided; as the objectives and length of the experiment. The participants were also informed that they could leave the experiment at any time as well as that participation was both voluntary and completely anonymous. Possible benefits and risks were emphasized; in addition, the contact details of the person in charge of the experiment were provided.

In addition to the information included in the document for the participants, a consent form was created (see Annexe I). This form contains information regarding the title of the experience, researcher and responsible faculty member. Participation was fully voluntary, and only proceeded with agreement of the terms and conditions of their involvement in the experiment.

This documentation was developed in line with the guidelines given by the *Commission d'éthique* at the University of Geneva that can be found in the document *Consignes pour la rédaction du formulaire d'information et de consentement*[8].

4.2.2 Instructions

As mentioned above, there were three sets of instructions required, namely one for the test group, a second for the control group and a third for the evaluators (see Annexes VII, VIII and IX). The only difference between the test and control groups' instructions was the use of PE guidelines. This is the reason why the first two sets of instructions were similar except for one extra step included for the test group, where the participants were asked to follow the PE guidelines (see Figure 16).

11. Comienza a poseditar la traducción propuesta siguiendo las directrices de posedición que encontrarás a continuación. También las encontrarás en un documento PDF individual en el paquete de la tarea.

	PE Guidelines for publishable quality
G1	Use as much of the raw MT output as possible, *but*:
G2	Ensure the target reader perfectly understands the content of the target text
G3	Ensure the target text communicates the same meaning and message as the source text
G4	Ensure target-text language is appropriate, but do not restructure sentences solely to improve the flow of the target text
G5	Aim for a grammatically, syntactically and semantically correct target text
G6	Ensure key terminology is correctly translated (considering context and the client's list of preferred terminology, if available)
G7	Ensure any untranslated terms belong to the client's list of 'Do Not Translate' terms, if available
G8	Edit any offensive, inappropriate or culturally unacceptable content for the target reader
G9	Apply basic rules regarding spelling and punctuation
G10	Ensure the same ST tags are present and in the correct positions in the target text

Figure 16: Extra step for the test group

The instructions documents followed the style used for describing the practical tasks in the courses of the Department of Translation Technology (referred to by its French acronym TIM). They were structured as follows: at the

[8] http://www.unige.ch/fapse/faculte/organisation/commissions/commission-ethique.html [Accessed 01/03/2015]

beginning of the first page information about the task was provided. The participants were also advised to read the instructions carefully before starting the task and to contact the researcher at any time if there were any doubts about the assignment. Both sets of instructions contained information about the particular task, namely of post-editing and human evaluation, as seen in Figure 17 below.

INSTRUCCIONES PARA LOS PARTICIPANTES

EVALUACIÓN HUMANA DE LA POSEDICIÓN

Estas instrucciones forman parte del proceso evaluación de la posedición humana realizada en la plataforma en línea MateCat.

Por favor, lee detenidamente estas instrucciones antes de comenzar con la tarea. Pregunta cualquier duda que se te plantee y/o comenta cualquier problema que se te presente.

Antes de comenzar haremos una pequeña introducción a la evaluación humana del proceso de posedición de un texto traducido por un sistema de traducción automática.

Existen dos formas de realizar esta evaluación: automática y humana. En este caso utilizaremos el segundo tipo.

Siguiendo a Hutchins y Somers, existen varias medidas para evaluar la calidad del texto: precisión (*accuracy*), claridad (*clarity*) y estilo (*style*).

También proponen analizar los errores. Para ello, el poseditor "[...] counts each addition or deletion for a word, each substitution of one Word by another, each instance of the transposition of words in phrases [...]" (Hutchins y Somers 1992, 164).

En nuestro caso mediremos la precisión de la versión poseditada con respecto al texto origen con una escala de 1 a 5, siendo 5 el nivel máximo de precisión. También anotaremos los errores siguiendo la definición del segundo método.

Figure 17: Instructions for evaluators

An additional instruction was included in the email sent to the participants. They were asked to note down the time they required to complete the task. This information was compared with the time provided by the Editing log of MateCat (see section 3.2.1).

4.2.3 Questionnaires

It was decided to create two types of questionnaire for each task: a general one that aimed to get general information about the participants, and a specific one related to the task in each case.

The use of an e-survey system like SurveyMonkey[9] or Zoomerang[10] was considered. This option was rejected, because these systems have a limited number of questions on their free account version. It was therefore decided to create the forms in Word documents and include them in the package created for each task.

As mentioned above, first the candidates had to read the information for participants' document and to complete the consent form. They could complete the general information questionnaire before or after finishing the task.

The general information questionnaire (see Annexe III) was the same for all the groups of participants. The goal of this set of questions was to obtain some general information about the participants and to facilitate the fulfilment of the questionnaires to the participants; it is easier to start answering general questions than beginning with specific ones. This questionnaire was divided into four parts. The first part aimed at collecting personal information about the participants (name, age, variant of Spanish, etc.). The second part included background questions about education, for instance, their last level of studies accomplished or their experience with English. The third and fourth parts contained questions about their experience and familiarity with computers in their everyday life, for example, what type of devices they have and use or which operating system(s) they usually use; the fourth part presented questions about participants' experience with CAT tools, specifically machine translation systems.

Three task-related questionnaires were designed (see Annexes IV, V and VI) with questions specific to each participant group, and a number of shared questions. All questionnaires were divided into two different sections. The first part contained questions about the task, for instance, the estimated time needed to complete the assignment. There were also questions related to machine

[9] https://www.surveymonkey.com/ [Accessed 17/03/15]
[10] http://www.zoomerang.com/ [Accessed 17/03/15]

translation, quality of the MT output, most common errors and which segments required most PE effort. The second part of the document was designed to obtain information about the post-editing experience of the candidates and their opinion regarding post-editing and the role of the post-editor. The post-editor questionnaires, included questions related to the use of PE guidelines.

All the questionnaires were created following a series of documentation elaborated by different universities (University of Colorado 1999, Stanford University 2011 and UC San Diego 2013). The methodologies proposed by Cornford and Smithson (2005), Oates (2006) and Jansen (2007) were also followed.

According to the documentation followed, the participants were informed that the fulfilment of the questionnaires was not compulsory but recommended for the evaluation. The length of the questionnaires was four pages maximum, with a reasonable number of 15/16 questions. The questions were organised from most general to most specific, avoiding non-specific questions. They were as short as possible avoiding slang, colloquial expressions and technical terms.

It was intended to use different types of questions: multiple choice, ranking and rating questions and also open-ended questions (Stanford University 2011). Both closed-ended and open-ended questions were used in order to obtain the appropriate information. Open-ended questions were carefully chosen in the section where the participants had to give their opinion about post-editing. Most of the other questions were closed-ended, since usually respondents prefer this type of questions because of their simplicity and the shorter time required to complete them, and their amenability to analysis.

4.2.4 Post-editing Evaluation Spreadsheet

An Excel spreadsheet model was required for the post-editing evaluation. Separated Excel spreadsheets for the post-editing evaluation, one for each post-edited version were created.

The different methods and aspects that were used to analyse the information obtained were described in section 3.3

The Excel model contained the following information:

Source Text	Original text in English
Raw MT Output	Text translated with MateCat
PE Version	Post-edited version of the pertinent participant
PE Accuracy 1=not accurate at all, 5= completely accurate	Measurement of the post-editing accuracy with a five degree scale
Number of modifications insertions | deletions | substitutions | shifts | others	Comparison of the raw MT output and the PE versions. Evaluators counted the number of modifications (insertions, deletions, substitutions and shifts) that the post-editors made while post-editing
Preferred version MT output | PE	Subjective opinion about the preferred version
Comments	Additional information

4.3 Pilot Test

It was decided to create a pilot test to test the materials created and the process proposed for the experiment.

The person chosen for the pilot test needed to have a profile as close as possible to that of the anticipated participants. She has Spanish as mother tongue, is a current student of the Master in Translation at the University of Geneva and has followed most of the courses of TIM.

This participant carried out the post-editing task with guidelines and the human evaluation. The reason why it was decided to perform this type of post-editing rather than the post-editing without guidelines was to test the additional steps that the participants had to perform. Additionally, the documentation was tested to find out possible errors as ambiguities, misspellings and misunderstandings.

4.3.1 Pilot Test Execution

A package with the materials and the required information was sent by email to the participant carrying out the pilot test.

The first phase of the pilot test was to proofread the different documentation to make sure that there were no linguistic or grammar errors and to review the questionnaires and instructions to find out possible problems as those defined in Oates (2006). It was important to know if the participants would have difficulties answering the questions, if the questions would be ambiguous or vague and if the pre-defined responses covered all potential answers. In addition, the participant was asked to give the estimated time required to perform the different tasks and to consider whether the instructions were understandable and easy to follow.

To finalise the pilot test, the participant was asked to perform the task of post-editing with guidelines, to carry out the human post-editing evaluation and to complete the questionnaires. The post-editing assignment was done with a different text than the text chosen for the official study in order to avoid the modification of the database of the program before starting the real evaluation.

The pilot test was very useful in strengthening the materials. The main suggestions were the addition of more screenshots and the reorganisation of some points of the instructions and screenshots so that the text would be easier to follow. The pilot participant also suggested reformulating one instruction. Regarding the questionnaires, in a number of cases the ordering of the questions was altered.

All the existing problems were evaluated and solved thanks to the comments and suggestions of the pilot test participant. In addition, it was possible to estimate the time required to perform the tasks.

4.4 Experiment Execution

The experiment execution took place after the pilot test, when the improvements suggested by the pilot test evaluator were applied.

We were aware of the importance of having all the participants doing the task in the same place and at the same time. Nevertheless, due to problems related to the availability of the participants, the experiment was carried out individually and remotely on the web. The deadline suggested was five days even though the time required to perform the task by the pilot test evaluator was about an hour and a half for the post-editing task and two hours for the post-editing evaluation. The participants could contact the researcher at any point if they needed some additional time in which to complete the task.

Participants received an email with some general instructions about the task and about the enclosed study material. In addition, they could find the name of the project that they had to use in each case as well as the username and password for the Gmail accounts that were created for them. The packages contained the specific material for each task and also some common documents:

- Consent form (Annexe I)
- Information for participants (Annexe II)
- General questionnaire (Annexe III)
- Specific questionnaire depending on the task: post-editing with guidelines (Annexe IV), post-editing without guidelines (Annexe V) and human PE evaluation (Annexe VI)
- Instructions depending on the task: post-editing with guidelines (Annexe VII), post-editing without guidelines (Annexe VIII) and human PE evaluation (Annexe IX)

For the first part of the study, two different packages were created for the two groups that would perform the post-editing task, namely the test and control groups. The original text in English was also included in the package.

The package for the test group, the group post-editing with PE guidelines, contained in addition a document with the guidelines in case the participants would like to have them separately

Once they had checked the materials included, they could start with the task. First, they had to go to the webpage of MateCat and log in with the details of the Gmail account included in the email. The following step was to create the project with the settings described in the instructions, in order to make sure that the user was connected with the correct details and the project created with the required settings. Once these steps were completed, participants were advised to read the user guide and/or watch the introductory video to get familiarised with the platform (see section 3.2.1).

Once familiarised with the tool, participants could start post-editing. They needed to be careful while selecting the source of the suggested translation that they had to post-edit; they always had to choose the translation proposed by the MT engine. This requirement was highlighted in the instructions due to its importance. Once this task was completed, participants were asked to complete the questionnaires.

During the post-editing evaluation, evaluators also received an email with information about the task and the materials included in the package. The package for evaluators contained two folders. The first contained instructions, questionnaires and information for participants' documentation; the second contained the Excel spreadsheet that they had to complete for each PE version (see Annexe IX). Additionally, the original text in English, a Word document with the raw MT output and also a document with the post-edited version corresponding to each case were also included following the suggestions of the pilot test participant, even though these documents were included in the Excel spreadsheet (see section 4.2.4).

The human evaluation was divided into two different tasks. First, they were going to work with different spreadsheets, one for each PE version. They had to rank in a five degree scale the accuracy of the post-edited version comparing it against the original text. The second part of the human PE evaluation was to compare the raw MT output against the post-edited version and count "[…] each addition or deletion for a word, each substitution of one Word by another, each instance of the transposition of words in phrases […]" (Hutchins and Somers 1992, 164). In addition, they were asked to choose the preferred version between the MT output and the PE version. To finalise the task, they had to complete the questionnaires.

The second part of the evaluation was to perform the revision step in MateCat (see section 3.2.1) in order to select subjectively the type of errors they found in the different segments. These results were compared with the results obtained after completing the Excel spreadsheets (see section 5.9)

4.5 Conclusion

This chapter has described the experiment design and its execution. Section 4.1, has explained how the experience was carried out during the development of this master's thesis was structured, e.g. types of assignments, division of the participants. Hereafter, section 4.2, described the creation of the different documentation. This section has been divided into three parts treating the different types of documents created, namely information for participants, instructions and questionnaires. The pilot test settings and execution as well as the description of the main experiment were described in sections 4.3 and 4.4 respectively.

5. Results and Analysis

This chapter describes the results obtained after the experiment and its analysis of the aspects evaluated as described in section 5.5.

The chapter is divided into the following sections. Section 5.1 presents the profile of the participants after the analysis of the general questionnaire. Sections 5.2 to section 5.9 present the results of the post-editing tasks and the human PE, combined with the results of the specific questionnaires.

The results of two participants out of eight were excluded. In one case MateCat did not save the results and in the second case, there was a problem while choosing the translation suggestion in MateCat. The participant did not follow the instructions where it was said that it was required to choose the MT suggestions for each segment. Nevertheless, the information given in the questionnaires by these participants was used.

5.1 Profile of the Participants

The first set of questions analysed was the general questionnaire (see section 4.2.3). These questions were created to have a precise idea of the profile of the participants and their knowledge about computers and CAT tools.

The questionnaire was divided into four parts, where the first two parts were addressed to obtain information about the participants. From these questions, the following information was extracted. All the participants had Spanish as native language and the variants of Spanish present in this study were from Spain and Argentina. The ages of the participants varied from 25-34; all had at least a university degree, and in most cases participants also had a master’s degree or were in the process of completing one. Most participants had more than ten years of experience studying English.

The third part of the questionnaire was related to participants’ experience with computers in general. As expected, all the participants were familiar with computers and owned a laptop. In most cases, they also owned a tablet. They were familiar with the most common operating systems (Windows and iOS) and used computers or similar devices every day.

The objective of the last part of the questionnaire was to establish their experience with CAT tools (see Figure 18). All the participants have worked at least with one CAT tool and most of them know and have used more than one. Regarding machine translation systems, the majority of them have used at least one machine translation platform online, the most popular is Google Translate. In some cases they also know more MT platforms. Only two of the participants know a post-editing tool.

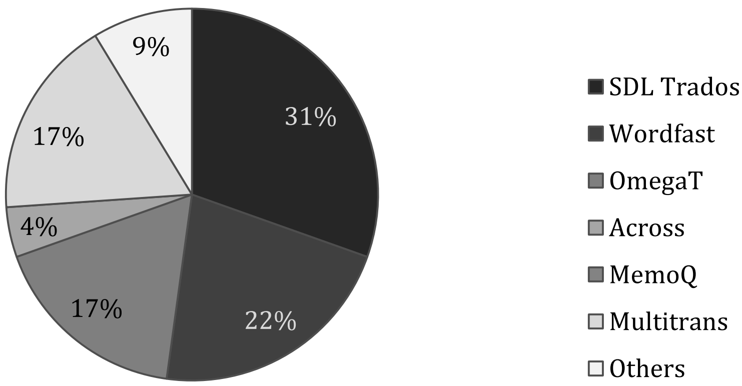


Figure 18: CAT tools used by the participants

5.2 Opinion about Post-editing

In the specific questionnaires (see section 4.2.3), participants answered questions about post-editing (questions nine, ten and eleven). In addition, they were asked about the importance of introducing a module about post-editing in the study program at the University.

All the participants thought that post-editing could be a useful practice, and agreed that post-editing could be very useful for translators because it could help translators to save time and effort while still obtaining a high quality output The main reason they gave was that translators will not have to translate the text from scratch or retranslate the text but instead focus on a reduced number of errors. However, even though the quality offered by the machine translation systems has improved greatly in recent years, and new platforms with new and more sophisticated features have been developed in the last decade, all participants agreed that the role of the human post-editor is, as yet, indispensable. They also commented that post-editing is maybe not suitable for all types of text and will be more suited to specific domains and situations, for example not very specialised texts.

Related to this point, participants reported a view that post-editing is not harmful for translators. Only one of the participants thought that this practice could be negative for translators. In addition, all agreed on the importance of introducing a module or at least an introduction to post-editing in the existing study program at the University. They think that probably the reason why some people are against post-editing is only because they do not have a lot of information about this field. They also pointed out the importance of the use and knowledge of CAT tools and the different roles of a translator nowadays. Translators should be aware of the possible roles that a translator can perform in order to form themselves accordingly.

5.3 Opinion about PE Guidelines

Both groups of post-editors were asked whether they consider the use of post-editing guidelines useful and/or necessary. The test group, which post-edited with guidelines (see section 4.2.2), was also asked whether the guidelines they had used were useful whether they were understandable, and whether participants would like to reformulate them. The control group, which worked without guidelines (see section 4.2.2), was asked if they would have liked to have some guidelines and if they think they would have made the same changes while post-editing if they had access to such guidelines.

All the participants stated that the use of guidelines while post-editing could be useful and that they are needed. The reasons they gave, in open-question number twelve, were that having a set of guidelines would reduce the number of changes and will save time and effort to the post-editor. The number and type of modifications would be limited, avoiding correcting more than necessary. Additionally, they felt that guidelines would help the post-editor avoid making subjective changes, meaning that the modifications should be similar between various post-editors working with the same set of guidelines.

The test group participants found the guidelines useful and understandable. They all agreed that the set of guidelines made the post-editing process easier and reduced the number of modifications. None felt that the guidelines needed a reformulation or any changes. However, two of the participants pointed out that it would be useful to have some examples of each rule to know exactly the type of changes required.

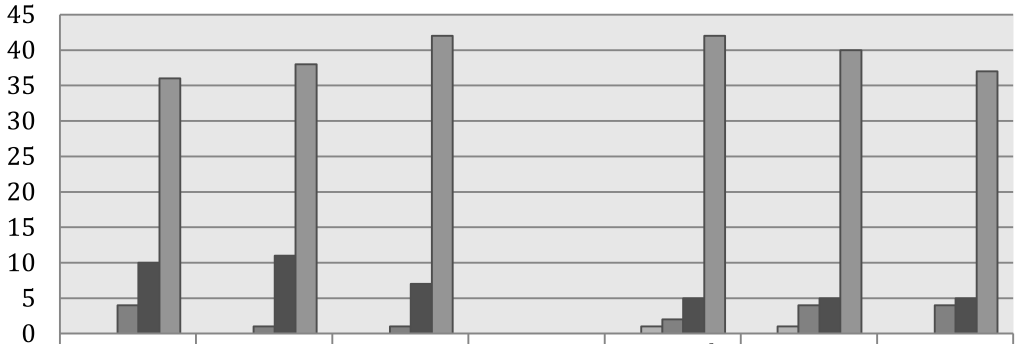
On the other hand, the control group concluded that they would have made different changes if a set of guidelines would have been available for them. They all mentioned that they made corrections related to the style but that were not necessary to understand the text. This type of change was only made to give a more natural aspect to the text but not because the text was incorrect.

5.4 PE Accuracy

An aspect analysed within the results of the human PE evaluation has been the post-editing accuracy.

In the Excel spreadsheet (see section 4.2.4) accuracy of the PE version against the original text was scored on a scale of five degrees, with one being the least accurate and five the most accurate. The two evaluators (see section 3.2.4) had to evaluate subjectively the 50 different segments that formed the text on this scale.

We analysed the results of both post-editing groups separately. The post-edited versions of the test group do not have any segment with the values number one or two. Two of the post-edited versions of the participants of the test group have a segment with an accuracy value number three. The third post-edited version had one had four cases with the value number three. The general accuracy of the three post-edited versions of the test group is between values four and five. In the case of the control group, we obtain similar results. One difference that could be found is the appearance of one segment in two post-edited versions with the value number two. There are also some cases with the value number three but, as mentioned, in general control group has a good level of accuracy. Figure 19 represents the five values of the scale and the number of segments corresponding to the each value.



	Test group				Control group		
1	0	0	0		0	0	0
2	0	0	0		1	1	0
3	4	1	1		2	4	4
4	10	11	7		5	5	5
5	36	38	42		42	40	37

Figure 19: PE Accuracy results per participant

Table 4 shows the average post-editing accuracy scores per group. The results concluded that, in general, the post-edited versions have a similar accuracy, having the majority of the segments the score of five. None of the segments had a "not accurate at all" value.

	Test Group					Control Group				
Accuracy values	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Average Accuracy per group	0	0	2	9.33	38.67	0	0.67	3.33	5	39.67

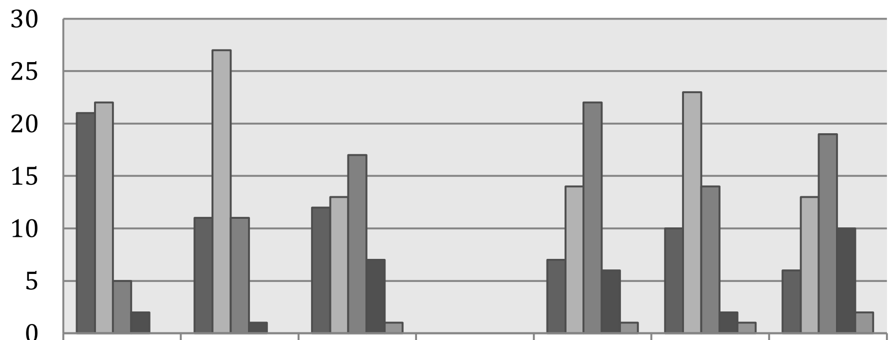
Table 4: Average PE Accuracy scores per group

5.5 PE Effort

With the results of the first part of the group-specific questionnaire and the information given by the editing log of MateCat (see section 3.2.1), the post-editing effort was analysed. Subjective effort was analysed by gathering information from the answers given by the participants in the questionnaire, while objective effort was analysed based on the information contained in the editing log and closer analysis of segments requiring a higher post-editing effort.

The MateCat editing log (see section 3.2.1) provides information about the post-editing effort required to edit the machine translation suggestions proposed by the platform. It is measured by the word edit distance between the first match provided and the final translation, and is given in percentages.

Figure 20 represents the results of MateCat editing log. The Figure shows the number of segments of the different post-edited versions and the associated PE effort. The results are presented separately for the test and control groups and organised depending on the PE effort given in percentage values. The results with 0% PE effort appear in the first column. The rest of the values are presented in different intervals that go from 0% to 45%. The last line shows the results higher than 45%. We did not want to have an extended number of intervals, we analysed the data provided by the editing logs from the different post-edited versions and realised that the values equal to 0% and higher than 45% and three intervals of 15% covered our data and our needs.



	Test Group				Control Group		
0%	21	11	12		7	10	6
0%-15%	22	27	13		14	23	13
15%-30%	5	11	17		22	14	19
30%-45%	2	1	7		6	2	10
>45%	0	0	1		1	1	2

Figure 20: MateCat Editing Log PE Effort

As it can be seen in Figure 20, the test group has more segments with 0% or with a PE effort between 0% and 15% than the control group. The post-edited versions of the test group have a lower number of segments with a PE effort between 30% and 45% than the control group. In addition, the test group has only one post-edited version with only one segment with a PE effort higher than 45% as opposite of the control group, there are four segments in total.

Table 5 presents the average PE effort per participant and per group according to the data gathered from the editing log provided by MateCat. The results show that the PE effort of the test group was lower for all the participants of this group. The results of the control group are clearly higher. Only the PE effort of one participant of the control group is similar to the results of the test group.

	Test group			Control group		
Average PE effort per participant	11%	6%	13%	19%	12%	21%
Average PE effort per group	10%			17.33%		

Table 5: Average PE Effort per participant and per group

Table 6 shows PE effort of the segment that required the highest effort to post-edit. The results of the participants of the control group are higher than the corresponding of the test group, being the highest percentage of all post-edited versions 76%.

Test group			Control group		
41%	32%	47%	76%	57%	54%

Table 6: Highest PE Effort per participant

Regarding the subjective results, the answers to the question number five of the specific questionnaire (see Annexes IV and V) were analysed. The participants were asked to comment what type of difficulties they consider required a higher post-editing effort. The answers of the candidates were similar, with none indicating a problem that was not raised by other participants. The segments that required a higher post-editing effort according to the participants were the segments that presented the issues in the following list:

- Specialised terminology: e.g. shock or export basket

Secs/Word	Words	Time-to-edit	PE Effort
19.4	21.00	06m:46s	24%

Segment	The link between macroeconomic volatility and the structure <g id="3">of a country's export basket has </g>been examined in a number of studies.
Suggestion	El vínculo entre la volatilidad macroeconómica y la estructura <g id="3">de la canasta de exportación de un país tiene </g>ha examinado en una serie de estudios.
Translation	El vínculo entre la volatilidad macroeconómica y la estructura <g id="3">de la canasta de exportaciones de un país </g>ha sido objeto de diversos estudios.
Diff View	El vínculo entre la volatilidad macroeconómica y la estructura <g id="3">de la canasta de ~~exportación~~ exportaciones de un país ~~tiene~~ </g>ha ~~examinado en una serie~~ sido objeto de diversos estudios.

- Complex sentences: e.g. subordinate clauses or several subjects

Secs/Word	Words	Time-to-edit	PE Effort
8	28.00	03m:42s	24%

Segment	Macroeconomic volatility is bad for development because it can reduce economic growth, make it difficult for households to smooth their consumption and adversely affect the distribution of income.
Suggestion	La volatilidad macroeconómica es malo para el desarrollo, ya que puede reducir el crecimiento económico, hacen que sea difícil para los hogares para suavizar su consumo y afectan negativamente a la distribución de la renta.
Translation	La volatilidad macroeconómica es negativa para el desarrollo ya que puede reducir el crecimiento económico, hacer que para los hogares sea difícil moderar el consumo y afectar negativamente a la distribución de la renta.
Diff View	La volatilidad macroeconómica es ~~malo~~ negativa para el ~~desarrollo,~~ desarrollo ya que puede reducir el crecimiento económico, ~~hacen~~ hacer que ~~sea difícil~~ para los hogares ~~para suavizar su~~ sea difícil moderar el consumo y ~~afectan~~ afectar negativamente a la distribución de la renta.

- Sentences with articles or relative pronouns missing

Secs/Word	Words	Time-to-edit	PE Effort
17.6	20.00	05m:51s	27%

Segment	For instance, supply-side constraints may make a country dependent on a narrow range of commodities for export and fiscal revenues.
Suggestion	Por ejemplo, limitaciones de la oferta pueden hacer depender un país en una estrecha gama de productos para la exportación y los ingresos fiscales.
Translation	Por ejemplo, las limitaciones de la oferta pueden hacer que un país dependa de una estrecha gama de productos básicos para obtener ingresos fiscales y de exportación.
Diff View	Por ejemplo, las limitaciones de la oferta pueden hacer ~~depender~~ que un país ~~en~~ dependa de una estrecha gama de productos básicos para ~~la exportación~~ obtener ingresos fiscales y ~~los ingresos fiscales.~~ de exportación.

We have analysed the editing log in order to find segments with the issues mentioned by the participants in the questionnaires. It can be seen that in the examples above, the PE effort required to post-edit the segments containing the main problems pointed out by the participants was higher than the average post-editing effort of both groups. As mentioned before, the average PE effort of the test group was 10% and the corresponding to the control group 17.33%.

5.6 Quality of the MT output

Post-editors also rated the general quality of the machine translation output with a scale of five values: very good, good, acceptable, bad and very bad. Table 7 presents the scores given by participants. The number indicates the number of participants that gave that value to the quality of the machine translation output. As can be seen, the quality of the output was considered as good and acceptable. None of the participants felt that the text was a bad translation and only one of them considered the MT output as a very good translation. It can be said that only a few major changes would be expected during the post-editing analysis.

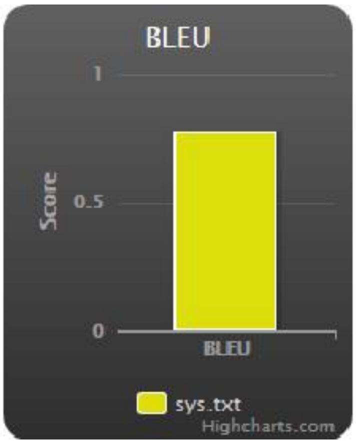
Very good	Good	Acceptable	Bad	Very bad
1	3	4	0	0

Table 7: Quality of the MT output according to the participants' opinion

In addition, the BLEU score was calculated to have an objective score of the quality of the raw MT output. BLEU is one of the most popular metrics to evaluate the quality of a machine translation output. BLEU score was calculated with Asiya

Online[11], an online platform where the user can obtain automatic evaluation scores from a selected set of metric representatives, e.g. BLEU, TER or METEOR. The usage of Asiya Online is intuitive and not very complicated, but the system is strict when accepting the inputs (original, reference and machine translated texts), in that the number of lines and segments have to be the same in all the different texts used. Due to this characteristic of the system and the fact that usually Spanish is a language that uses more words than English, we had to count the lines and segments in every case we performed an automatic evaluation within this platform to make sure that the numbers coincided.

The BLEU score obtained when evaluating the raw MT output was 0.77 (see Figure 21), which means that the MT output has a quality between medium and good. If we compare the results from the questionnaire and the BLEU score, it can be said that the opinions of the participants and the automatic evaluation are not very different. As mentioned, the majority of the participants considered that the quality of the text was acceptable and good.



BLEU	GTM-3	NIST	-WER	-PER	Ol	-TERbase	METEOR-ex	ROUGE-L
0.7754	0.6477	4.2379	-0.1429	-0.1071	0.8065	-0.1429	0.5354	0.8889

Figure 21: BLEU score

5.7 Time

In this section, the analysis of time required to perform the task is presented. The information was obtained from two different sources. First, we

[11] http://asiya.cs.upc.edu/demo/asiya_online.php [Accessed 10/07/2015]

analysed the time taken by the participants in the questionnaire. Second, the information provided in MateCat editing log was studied.

In the first part of the specific questionnaire (see section 4.2.3) the participants recorded the approximate time they needed to perform the complete task. Some of the participants performed the task over several days depending on their availability. In general, the participants of the control group needed more time to perform the task; for this reason it could be estimated that the post-editing effort was higher in the case of the group that worked without PE guidelines. Nevertheless, these results will be compared with the information provided by MateCat because the time given by the participants includes the reading of the instructions, getting familiarised with the tool and que fulfilment of the questionnaires.

In Table 8, the subjective time the participants needed to complete the task is shown.

In addition, the average time per group has been calculated. The participants of the test group required an average time of 02h20m to perform the task, while the control group needed 3h.

	Test group			**Control group**		
Time per participant	02h	02h	03h	03h	03h	03h
Average time per group	02h20m			03h		

Table 8: Subjective time needed per participant and average time per group

We then examined data gathered in the MateCat editing log to analyse the time required by participants to perform the post-editing task, excluding the time required to get familiar with the platform and to complete the questionnaires.

This information is shown in Table 9, where it can be seen that the participants of the control group required slightly more time to post-edit the text. The test group participants, who required an average time of 43m13s to perform the post-editing task, were faster than the control group participants, who required an average of 01h20m.

	Test Group			Control Group		
Time per participant	37m30s	43m49s	47m40s	01h49m42s	01h09m55s	01h04m03s
Average time per group	43m13s			01h20m		

Table 9: Time taken to complete the post-editing task, based on MateCat data, per participant and average time per group

5.8 Evaluation of the PE outputs

In this section the quality of the different post-edited versions is evaluated. On the one hand, the results of one section of the Excel spreadsheet (see section 4.2.4), where the evaluators counted the minimum number of modifications in every segment, in order to have the different types and number of modifications made during the post-editing task will be analysed. It had been intended to use HTER scores (see section 2.8.1) calculated by MateCat. However, when the results were exported, all the percentages were equal to 0%. Due to this technical issue in the system, it was decided to calculate the TER score (see section 2.8.1) within the platform Asiya Online[12] in order to have an objective result of the editing steps (see section 2.8.1) and not only the subjective opinion from the participants. We performed a TER evaluation for each post-edited version, comparing the original text, the post-edited version and the raw MT output. As mentioned in section 5.6, the Asiya platform is very strict with the number of lines and segments of the inputs, which is why the performance of the TER scores was a time-consuming task.

Table 10 presents the TER scores obtained after running the automatic evaluation, and indicate differences between the two groups. The test group TER scores (on average -0.18) were lower than those of the control group (on average -0.33), reflecting a lower number of editing steps, and more homogeneous.

[12] http://asiya.cs.upc.edu/demo/asiya_online.php [Accessed 10/07/2015]

	Test Group			Control Group		
TER	-0.14	-0.21	-0.20	-0.43	-0.23	-0.33
Average TER per group	-0.18			-0.33		

Table 10: TER scores per participant and average TER score per group

After analyzing the TER scores, they showed that a difference between the test and the control groups' scores existed. The means of the two groups having and average TER score of -0.18 and -0.33 respectively seemed to differ. We analyzed whether the differences were explained by randomness or whether they were statistically significant, i.e. whether TER was affected negatively by using guidelines or not. In order to do so we applied a *t*-Test[13] that allows a comparisson of two averages. We were interested in whether using guidelines improve the TER score or not by testing if the sample mean of the test group was smaller than the control group sample mean, which made up the hypothesis we wanted to test. We wanted to prove that using guidelines made our post-editors not less effective than without guidelines. This leaded to a one-tailed t-test to be the tool of choice. The null hypothesis was then defined as there were no true differences between the score means, i.e. the difference corresponded to zero. The significance is expressed by *p*-values, which is a value between 0 and 1, interpreted as follows:

- p-value ≤ 0.05: The difference between the results is significant. The null hypothesis is rejected.
- p-value > 0.05: The results are not consistent. The evidence against the null hypothesis is weak and hence the null hypothesis still holds.

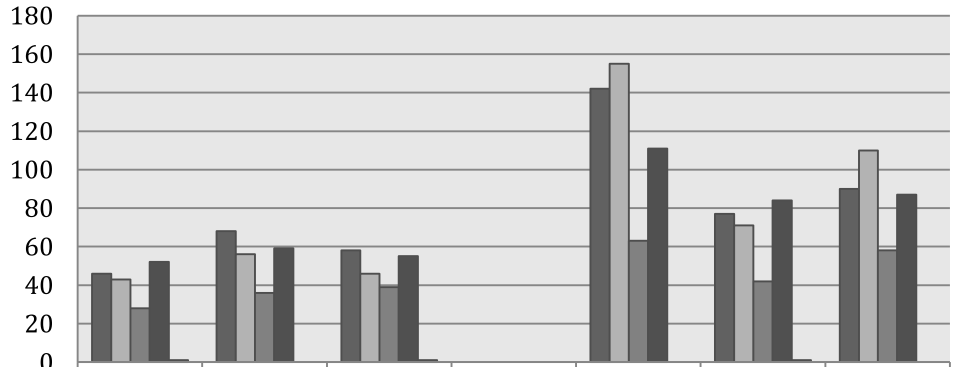
The *p*-value is essentially a probability. It expresses the likelihood how plausible it is that there is no difference between the TER scores of the two groups. In our case, the *p*-value computed corresponded to 0.048, the alternative hypothesis was true and the evidence was significant. According to this result, the

[13] http://archive.bio.ed.ac.uk/jdeacon/statistics/tress4a.html#Student%27s%20t-test [Accessed 06/08/15]

use of PE guidelines had a proven positive impact in the quality of the post-edited versions.

Figure 22 presents the number and types of modifications counted manually during the human post-editing evaluation. The types of modifications analysed were insertions, deletions, substitutions and shifts. Different types of modifications were counted under the column “others”, for example the change from brackets to square brackets in the case of two participants.

All the participants of the test group made fewer modifications in all categories than the participants of the control group. In addition, it can be seen that the number of modifications of the test group are quite similar, in contrast to the number of modifications made by the control group. The results of the control group are more variable, especially the first participant of the control group, who seemed to be unusually active. This might reflect the subjectivity of the participants that post-edited the text without guidelines.



	Test Group				Control Group		
Insertions	46	68	58		142	77	90
Deletions	43	56	46		155	71	110
Substitutions	28	36	39		63	42	58
Shifts	52	59	55		111	84	87
Others	1	0	1		0	1	0

Figure 22: Number and types of modifications according to the human evaluation

The automatic and human evaluation produced similar results, with the test group scores lower and more homogenous than those of the control group. As mentioned, we think that this may be caused because the participants of the control group were more subjective in the absence of PE guidelines when post-editing: they did not have any set of guidelines to follow, so they had complete freedom to modify the text.

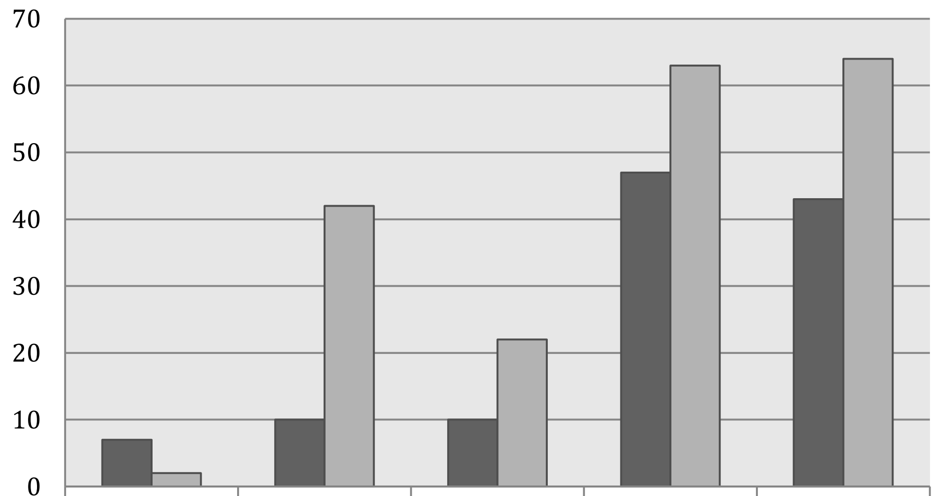
5.9 Types of errors in the MT output

Participants also identified the error types occurring most commonly in machine translation output (see section 4.2.3). In addition, the evaluators were asked to select the type of errors existing in each segment of the text during the revision step in MateCat (see section 3.2.1).

The errors that participants considered as occurring most frequently were:

- Mistranslations
- Punctuation
- Grammar: concordance, omission of articles.
- Style: e.g. sentences with an English structure, not incorrect but not natural in Spanish.

During the revision phase, the evaluators were able to select from a list (see section 3.2.1) the types of errors they found in the different segments. The more frequent issues in both groups were translation errors, and issues of quality of the language and style. They coincide with the issues indicated by the post-editors in the questionnaires. As it can be seen in Figure 23, the post-edited versions of the control contain a higher number of problems, especially issues regarding language quality and style. A possible conclusion can be that the test group gave less attention to errors that were not fundamental to understand the text, probably because such errors were not prioritised within the guidelines if the text was understandable.



	Tag issues	Translation errors	Terminology	Language Quality	Style
Test group	7	10	10	47	43
Control group	2	42	22	63	64

Figure 23: Types of errors selected by the participants during the revision phase

5.10 Conclusion

In this chapter the results for the post-editing experiment, the human post-editing evaluation and the questionnaires have been analysed.

In general, the results show a positive impact of the use of PE guidelines while post-editing a raw MT output. Additionally, the participants reported that the use of PE guidelines can be useful for translators. Table 11 below summarises the results obtained after the analysis.

VARIABLES EVALUATED	OBJECTIVE EVALUATION	SUBJECTIVE EVALUATION
Opinion about PE	-	Useful and not harmful for translators: PE helps to reduce time and effort
Opinion about PE guidelines	-	Useful and needed to PE: guidelines help to reduce the number of modifications and save time and effort
PE Accuracy	-	Similar accuracy in both groups. The majority of the segments have accuracy values between 4 and 5
PE effort	Test group: - 11% - 6% - 13% Average PE effort: 10%	All the participants considered that the segments with the following issues required a higher PE effort: - Specialised terminology

<table>
<tr>
<td></td>
<td>Control group:
- 19%
- 12%
- 21%
Average PE effort: 17.33%</td>
<td>- Complex sentences
- Sentences with articles or relative pronouns missing</td>
</tr>
<tr>
<td>Quality of the MT output</td>
<td>BLEU score: 0.77</td>
<td>Very good: 1 participant
Good: 3 participants
Acceptable: 4 participants
Bad: 0 participants
Very bad: 0 participants</td>
</tr>
<tr>
<td>Time</td>
<td>Test group:
- 37m30s
- 43m49s
- 47m40s
Average time: 43m13s

Control group:
- 01h49m42s
- 01h09m55s
- 01h04m03s
Average time: 01h20m</td>
<td>Test group:
- 02h
- 02h
- 03h
Average time: 02h20m

Control group
- 03h
- 03h
- 03h
Average time: 03h</td>
</tr>
<tr>
<td>Evaluation of the PE outputs</td>
<td>Test group:
- -0.14
- -0.21
- -0.20
Average TER: -0.18

Control group:
- -0.43
- -0.23
- -0.33
Average TER: -0.33

TER t-Test: 0.048</td>
<td>
<table>
<tr><td></td><td colspan="3">Test group</td></tr>
<tr><td>Insertions</td><td>46</td><td>68</td><td>58</td></tr>
<tr><td>Deletions</td><td>43</td><td>56</td><td>46</td></tr>
<tr><td>Substitutions</td><td>28</td><td>36</td><td>39</td></tr>
<tr><td>Shifts</td><td>52</td><td>59</td><td>55</td></tr>
<tr><td>Others</td><td>1</td><td>0</td><td>1</td></tr>
</table>
<table>
<tr><td></td><td colspan="3">Control group</td></tr>
<tr><td>Insertions</td><td>142</td><td>77</td><td>90</td></tr>
<tr><td>Deletions</td><td>155</td><td>71</td><td>110</td></tr>
<tr><td>Substitutions</td><td>63</td><td>42</td><td>58</td></tr>
<tr><td>Shifts</td><td>111</td><td>84</td><td>87</td></tr>
<tr><td>Others</td><td>0</td><td>1</td><td>0</td></tr>
</table>
</td>
</tr>
<tr>
<td>Types of errors in the MT output</td>
<td>Test group:
- Tag issues: 7
- Translation errors: 10
- Terminology: 10
- Language Quality: 47
- Style: 43

Control group:
- Tag issues: 2
- Translation errors: 42
- Terminology: 22
- Language Quality: 63
- Style: 64</td>
<td>All the participants considered the following issues as the more frequent in the MT output:

- Mistranslations
- Punctuation
- Grammar
- Style</td>
</tr>
</table>

Table 11: Summary of variables evaluated and results

Conclusion

The objective of this master's thesis was to analyse the impact of the use of a specific set of guidelines while post-editing a raw machine translation output.

First, the basic concepts of machine translation and post-editing were described in order to provide a theoretical foundation for the subsequent analyses and experiments.

The experiment developed was divided into three steps. The first was a pilot test that was carried out by one participant. It helped us to test the experiment design and the materials created.

During the second step, post-editing tasks were performed by selected participants that meet the following requirements: Spanish native language students or recent graduates from the FTI that had undertaken at least one TIM course. Participants post-edited an extract of the World Trade Report 2014 of the WTO that had been translated with MateCat, an online MT platform. The participants were divided into two groups: one test and one control group. Both groups of participants had to post-edit the same raw MT output, the test group with PE guidelines and the control group without guidelines. In addition, all participants completed two sets of questionnaires in order to obtain more information about the profile of the participants and about the specific task they performed and to know their opinion about post-editing.

The final phase of the experiment was carried out by two evaluators, who performed the human post-editing evaluation. They were also Translation and Interpreting graduates with Spanish as native language.

At a first glance, different conclusions can be drawn from the analysis. In general, it can be concluded that the participants find the use of guidelines useful and that the use of post-editing guidelines while post-editing a raw MT output appears to have a positive impact in terms of time required, post-editing effort and accuracy. Both groups of participants produced work of similar quality, but the participants of the test group were quicker and more consistent.

All the participants of the test group performed the post-editing task in less time than the participants from the control group. None of the participants of the test group required an hour to post-edit the text, with an average time of 43m13s. All the participants of the control group needed more than an hour with an average time of 01h20m. Regarding the PE effort, both groups of post-editors pointed out the same issues that leaded to a higher PE effort. However, the post-editing effort as obtained from the editing log of both groups indicated that the PE effort was higher for the control group (17.33%) than for the test group (10%). Our results reflect that time and post-editing effort are strongly associated. However, we are aware of the limit number of participants in our study to state a correlation. It would be interesting to reproduce the experiment in a larger scale and analyse the results.

In addition, the participants rated the quality of the MT output similarly; seven out of eight participants considered the quality of the MT output between "good" and "acceptable" on a five-point scale.

For this particular experiment design, greater effort appears to be associated with greater time spent. Concluding from this study, the post-editing effort is higher in the cases were the participants required more time to perform the task. The duration of the post-editing process was shown to be longer when the post-editing effort is also high.

The subjective results of the quality evaluation of the post-editing outputs present bigger differences than the objective results. Nevertheless, the final results are similar and the participants that made more post-editing changes according to the human evaluation also got higher TER scores. According to the *t*-Test carried out, the TER scores are statistically significant with a *p*-value of 0.048.

The types of errors found in the MT output by the evaluators and the ones pointed out by the participants in the questionnaire relate to some extent. The issues categories of "tag issues" and "terminology", present in MateCat, were not pointed out by the participants. This was probably because the participants noted the terminology issues as the problems of the MT output and not when commenting the issues related to the post-edited version.

In this study, we used the set of post-editing guidelines proposed by Flanagan and Paulsen (2014), which were an improved version of the TAUS guidelines for publishable quality. Our participants found these guidelines to be clear, understandable, and useful for post-editing a raw MT output: no major problems or issues with the guidelines were reported.

To conclude after this brief overview of the results obtained after the experiment, we can say that there are differences between the performances of the two groups that participated in this experience. However, even if there are no huge differences, we can say that the use of post-editing guidelines could have a positive impact reducing the PE effort and the time required to post-edit. As suggested by Allen (2003), the use of PE guidelines would also avoid a certain degree of subjectivity when post-editing, which will help to have texts with a similar quality and homogeneity between a group of post-editors. In addition, the set of guidelines proposed by Flanagan and Paulsen (Flanagan and Paulsen 2014) has proven to be useful and clear.

Limitations and Further Work

This study has several limitations. The number of participants was limited. As we decided to come up with two groups for the post-editing task and one for the evaluation, we divided the number of participants in three different groups. In addition, the results of two prospective participants were invalid which further reduced the total number of cases available for the study. One way to overcome this issue could have been to create only two groups of participants and exchange the results between the groups for the evaluation step. Another possibility to proceed with the experiment may have been to switch tasks between participants having guidelines and not, to examine whether the contrasts between test and control groups relate to the people doing the tasks. Additionally, only one domain in only one pair of languages has been tested. We suggest future studies to focus on multiple domains and language pairs in order to further generalize the results.

The text chosen was also a limitation. We thought that it was possibly less likely to get participation if the task was more demanding, for this reason the length was reduced to 50 segments, we are aware that it is not a high number of segments and that the results would be more reliable with a larger number of segments.

Another limitation is the machine translation platform used. Nowadays, there are several online platforms and commercial systems in the market, which made for a difficult choice. We decided to choose MateCat because it is a platform that had not been used before for a master's thesis at the University of Geneva, and it is not known widely amongst students and translators. It would be interesting to use a different platform to know to what extend the results would vary.

Since our study has been conducted on a small scale, our motivation was to identify starting points for future large scale studies. It would be interesting for similar work in the future to use lists of vocabulary of a specific domain and/or a translation memory to analyse whether the results would reinforce ours in terms of quality and time. Probably, having lists of vocabulary and/or translation memories would reduce the time required and would improve the quality of the post-edited versions giving more homogeneity to the terminology. Another possible and interesting topic for a future study would be to perform a similar experiment but to also analyse the "good enough" quality: in this study only the post-editing publishable quality has been covered. This way various degrees of PE efforts could be analysed relative to the translation quality required. Finally, a further potential topic of study would be to compare across a range of different guidelines, for example GALE, and MT systems to know which combination could be more efficient and useful.

Bibliography

ALLEN Jeffrey (2003) "Post-Editing". *In:* SOMERS H., *Computers and Translation: A Translator's Guide*. Amsterdam-Philadelphia: John Benjamins, pp. 297-318.

ARNOLD Douglas et al. (1994) *Machine translation: an Introductory Guide*. London: NCC Blackwell.

BOUILLON Pierrette (2012) *Traduction Automatique 1*. Geneva: University of Geneva, not published.

BOUILLON Pierrette (2013) *Traduction Automatique 2*. Geneva: University of Geneva, not published.

CORNFORD, T. and SMITHSON, S. (2005) *Project Research in Information Systems: A Student's Guide*. New York: Palgrave Macmillan.

FERNANDEZ GUERRA Ana (2000) *Machine translation: capabilities and limitations*. Valencia: Artes Gráficas Soler.

FLANAGAN Marian and PAULSEN Tina (2014) "Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes", *The Interpreter and Translator Trainer*, Vol. 8, No. 2, 257-275. Available at: http://www.tandfonline.com/doi/full/10.1080/1750399X.2014.936111 [Accessed 12/2014].

GIACHETTI Mary-Sue (2013) *L'Intégration de la traduction automatique dans le processus de localisation à Autodesk: Étude de cas et enquête auprès des traducteurs*. Geneva: University of Geneva.

GRAY Charlotte (2014) *A Comparative Evaluation of Localisation Tools: Reverso Localize and SYSTRANLinks.* Geneva: University of Geneva.

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*, Vol. 7, Issue 1.

GUZMÁN Rafael (2007) "Manual MT Post-editing: if it's not broken, don't fix it!", *Translation Journal*, Vol. 11, No. 4. Available at: http://www.translationjournal.net/journal/42mt.htm [Accessed 12/2014].

HUTCHINS John W. and SOMERS Harold L. (1992) *An Introduction to machine translation*. London: Academic Press.

HUTCHINS John W. (1994) "Machine translation: history and general principles". *In*: *The encyclopaedia of languages and linguistics*. Oxford: Pergamon Press, vol. 5, pp. 2322-2332.

HUTCHINS John W. (2006) "Machine translation: a concise history". Available at: http://www.hutchinsweb.me.uk/CUHK-2006.pdf [Accessed 18/05/2015].

JANSEN, Karen J., CORLEY Kevin G., JANSEN Bernard J. (2007) "E-Survey Methodology", Chapter 1. Available at: http://faculty.ist.psu.edu/jjansen/academic/pubs/esurvey_chapter_jansen.pdf [Accessed 23/03/2015].

JURAFSKY Daniel and MARTIN James H. (2000) *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* New Jersey: Prentice Hall.

KOEHN Philipp (2007) "Statistical Machine Translation – Tutorial". University of Edinburgh. Available at: http://www.mt-archive.info/MTS-2007-Koehn-3.pdf [Accessed 21/05/2015].

KOEHN Philipp (2010) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

KRINGS Hans P. and KOBY Geoffrey S. (2001) *Repairing Texts. Empirical Investigations of Machine Translation Post-Editing Processes*. Kent: The Kent State University Press.

LEVENSHTEIN, Vladimir I. (1965) *Binary codes capable of correcting spurious insertions and deletions of ones. Problems of Information Transmission* vol. 1, pp. 8–17.

L'HOMME Marie-Claude (2008) *Initiation à la traductique*. Montréal: Linguatech.

MATECAT (2014) *Website*. Available at: https://www.matecat.com/ [Accessed 06/2015).

MOSES – STATISTICAL MACHINE TRANSLATION SYSTEM (2013) *Website*. Available at: http://www.statmt.org/moses/ [Accessed 21/05/2015].

NIST (2009) "NIST Machine Translation Evaluation for GALE". Available at: http://www.itl.nist.gov/iad/mig/tests/gale/index.html [Accessed 17/02/2015].

OATES Briony J. (2006) *Researching Information Systems and Computing*. London: SAGE Publications Ltd.

O'BRIEN Sharon (2002) "Teaching Post-editing: A Proposal for course Content". Available at: http://mt-archive.info/EAMT-2002-OBrien.pdf [Accessed 22/05/2015].

O'BRIEN Sharon (2004) "Machine Translatability and Post-Editing Effort: How do they Relate?", *Translating and the Computer*, nº26. Available at: http://mt-archive.info/Aslib-2004-OBrien.pdf [Accessed 11/2014].

O'BRIEN Sharon (2011) "Towards Predicting Post-Editing Productivity", *Machine Translation*, Vol. 25, nº1, pp. 197-225.

O'BRIEN Sharon (2012) "Translation as a human-computer interaction", *Translation Spaces*, nº1, pp. 101-122.

PERON Cristina (2013) *Évaluation d'une plate-forme de localisation: le cas de Reverso Localize.* Geneva : University of Geneva.

PLITT Mirko and MASSELOT François (2010) "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context", *The Prague Bulletin of Mathematical Linguistics*, nº93, pp. 7-16. Available at: http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf [Accessed 11/2014].

REVERSO-SOFTISSIMO (2013) *Website*. Available at: http://localize.reverso.net [Accessed 20/05/2015].

RICO Celia and TORREJÓN Enrique (2012) "Skills and profile of the new role of the translator as MT post-editor", *Revista Tradumàtica: tecnologies de la traducció*, nº10, pp. 166-178. Available at: http://ddd.uab.cat/pub/tradumatica/tradumatica_a2012n10/tradumatica_a2012n10p166.pdf [Accessed 15/01/2014].

ROBERT Anne-Marie (2010) "La post-édition: l'avenir incontournable du traducteur?", *Traduire*, nº222. Available at: http://tiltcommunications.fr/imges/Post_edition_AMR_2010_Final.pdf[Accessed 25/01/2015].

ROBERT Anne-Marie (2013) "Vous avez dit post-éditrice? Quelques éléments d'un parcours personnel", *The Journal of Specialised Translation*, nº19. Available at: http://www.jostrans.org/issue19/art_robert.php [Accessed 20/01/2015].

SCHÄFER Falko (2003) "MT post-editing: How to shed light on the 'unknown task' – Experiences made at SAP" [online]. Walldorf: SAP. Available at: http://mt-archive.info/CLT-2003-Schaefer.pdf [Accessed 10/2014].

SDL (2014) *Website*. Available at: http://www.sdl.com [Accessed 19/05/2015].

STANFORD UNIVERSITY (2011) "Tips for Survey Design". Standford: Standford University. Available at: http://web.stanford.edu/group/ssds/cgi-bin/drupal/files/Guides/Tips_for_survey_design_2011.pdf [Accessed 23/03/2015].

TAUS (2010) "MT Post-editing Guidelines". Available at: https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines [Accessed 15/05/2015].

THICKE Lori (2013) "Post-Editing Best Practices Identified by ACCEPT" [Blog] *LexWorks: Machine Translation that Works*. Available at: http://lexworks.com/translation-blog/3-sources-for-post-editing-best-practices [Accessed 15/02/2015].

TRUJILLO Arturo (1999) *Translation Engines: Techniques for Machine Translation*. London: Springer-Verlag.

UC SAN DIEGO (2013) "Writing Good Survey Questions – Tips & Advice". San Diego: Student Research & Information UC San Diego. Available at: http://studentresearch.ucsd.edu/_files/assessment/workshops/2013-02-21_Writing-Good-Survey-Questions-Handouts.pdf [Accessed 23/03/2015].

UNIVERSITY OF COLORADO (1999) "Suggestions for Wording Survey Questions". Boulder: Department of Communication University of Colorado. Available at: http://comm.colorado.edu/~freyl/courses/quantitative-communication-research-methods/exercises-and-handouts/sampling-&-survey/survey-question-wording.pdf [Accessed 23/03/2015].

VASCONCELLOS Muriel and LÉON Marjorie (1985) "SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization", *Computational Linguistics*, nº11, pp. 122-136.

WAGNER Emma (1985) "Post-Editing Systran – A Challenge for Commission Translator", *Terminiologie et Traduction*, nº3, pp. 1-7. Available at: http://mt-archive.info/T&T-1985-Wagner.pdf [Accessed 25/01/2015)

Annexes

Annexe I

UNIVERSITÉ DE GENÈVE

FORMULARIO DE CONSENTIMIENTO

TÍTULO DEL EXPERMIENTO
Analysis of the Impact of the Use of Post-editing Guidelines in a Raw Machine Translation Output

PERSONA A CARGO DE LA INVESTIGACIÓN
Remedios Hidalgo Ríos, estudiante de máster en Traducción en el Dpto. de Tratamiento Informático Multilingüe.

MIEMBRO RESPONSIBLE DE LA FACULTAD
Prof. Pierrette Bouillon

DECLARACIÓN
Acepto y estoy de acuerdo con participar en este estudio:

- He leído la información proporcionada y entiendo los objetivos del estudio.
- He obtenido una respuesta satisfactoria por parte de la persona a cargo de la investigación a las preguntas que he realizado.
- Entiendo los términos de mi participación en este estudio.
- Entiendo que los resultados del estudio llevado a cabo con esta experiencia serán utilizados por la persona a cargo de la investigación, que mi participación es voluntaria, y que puedo abandonar en cualquier momento.

NOMBRE DEL PARTICIPANTE

FIRMA DEL PARTICIPANTE

FIRMA DEL INVESTIGADOR

FECHA

Annexe II

UNIVERSITÉ DE GENÈVE

INFORMACIÓN PARA LOS PARTICIPANTES

Analysis of the Impact of the Use of Post-editing Guidelines in a Raw Machine Translation Output

OBJETIVOS DEL EXPERIMENTO

Este experimento forma parte de una tesina de máster. El objetivo es analizar el impacto del uso de directrices de posedición durante la posedición de textos traducidos con un sistema de traducción automática.

La tarea de los participantes será de llevar a cabo una serie de tareas de posedición y evaluación, así como completar dos tipos de cuestionarios, uno de información general y otro con relativo a la tarea realizada y sus resultados.

DURACIÓN DEL EXPERIMENTO

El plazo para realizar la tarea será de 5 días a contar desde el día de envío y recepción de las instrucciones. Se aconseja realizar la tarea en un mismo día, no obstante no es un requisito obligatorio.

Podrán realizarse de manera autónoma o con presencia del investigador.

ABANDONO DE LA EXPERIENCIA

Podrás abandonar la experiencia en cualquier momento y sin ningún tipo de consecuencia.

PARTICIPACIÓN VOLUNTARIA

La participación en esta experiencia es completamente voluntaria.

BENEFICIOS Y RIESGOS

Durante esta experiencia llevarás a cabo la tarea de posedición en dos situaciones diferentes, con y sin directrices de posedición; así como evaluación de la posedición. Además, tendrás la oportunidad de reflexionar sobre las ventajas e inconvenientes de la posedición como posible trabajo para los traductores.

La participación en esta experiencia no conlleva ningún riesgo. Si en cualquier momento no estás de acuerdo con el proceso, puedes abandonar sin ningún problema.

PROTECCIÓN DE LA INFORMACIÓN

Deberás proporcionar información personal como tu nombre, edad o nivel de estudios. Tu perfil de participante será asociado a un número de usuario durante la experiencia. Esta información no se compartirá, así como los resultados que se extraigan serán anónimos en la tesina.

INFORMACIÓN DE CONTACTO

Para cualquier tipo de información, por favor contacta con Remedios Hidalgo a través de la siguiente dirección de correo: remedios.hidalgo@etu.unige.ch

Annexe III

Usuario nº:

UNIVERSITÉ DE GENÈVE

CUESTIONARIO PARA PARTICIPANTES

INFORMACIÓN GENERAL

Rellena el cuestionario que se presenta a continuación. Las preguntas con un asterisco (*) al lado del enunciado pueden tener más de una respuesta.

SECCIÓN 1: INFORMACIÓN PERSONAL

Nombre: ____

Apellido(s): ____

Género

☐ Masculino

☐ Femenino

Edad ______________________________

Lengua(s) materna ______________________________

Variante(s) del español *

____

¿Vives en un país hispanohablante actualmente?

☐ Sí

☐ No

Si la respuesta es "no" : ¿en qué país? ¿cuánto tiempo?

____

SECCIÓN 2: EDUCACIÓN

Último nivel de estudios completado

☐ Ninguno
☐ Educación primaria
☐ Educación secundaria
☐ Bachillerato
☐ Diplomatura/Licenciatura/Grado
☐ Máster
☐ Doctorado
☐ Otros ____________________________________

Formación en curso

☐ Ninguno
☐ Educación primaria
☐ Educación secundaria
☐ Bachillerato
☐ Diplomatura/Licenciatura/Grado
☐ Máster
☐ Doctorado
☐ Otros ____________________________________

Conocimiento de lengua inglesa

☐ Ninguno
☐ Menos de 2 años de estudio
☐ Entre 2 y 5 años de estudio
☐ Entre 5 y 10 años de estudio
☐ Más de 10 años de estudio
☐ Bilingüe
☐ Lengua materna
☐ Otros ____________________________________

SECCIÓN 3: CONOCIMIENTOS INFORMÁTICOS

¿Tienes ordenador o algún otro tipo de dispositivo electrónico?

☐ Sí

☐ No

¿De qué tipo? *

☐ Ordenador de mesa

☐ Ordenador portátil

☐ Tableta

☐ Otros ____

¿Utilizas algún dispositivo electrónico a diario?

☐ Sí

☐ No

¿Cuántas horas al día?

☐ Menos de una hora al día

☐ 1 hora al día

☐ 2 horas al día

☐ 3 horas al día

☐ 4 horas al día

☐ Más de 4 horas al día

¿Para qué tareas? *

☐ Trabajo

☐ Estudios

☐ Consulta

☐ Ocio

☐ Otros ____

¿Qué sistema(s) operativo(s) conoces? *

☐ Windows

☐ iOS

☐ Linux

☐ Ubuntu

☐ Android

☐ Otros ______________________________

¿Qué sistema(s) operativo(s) utilizas normalmente? *

☐ Windows

☐ iOS

☐ Linux

☐ Ubuntu

☐ Android

☐ Otros ______________________________

SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

¿Has trabajado con alguna herramienta de TAO?

☐ Sí

☐ No

¿Con cuál(es)? *

☐ SDL Trados

☐ Wordfast

☐ OmegaT

☐ Across

☐ MemoQ

☐ Multitrans

☐ Otros ______________________________

¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? *

☐ Experto: ____

☐ Avanzado: ____

☐ Medio: ____

☐ Principiante avanzado: _______________________________

☐ Principiante: _______________________________________

¿Sueles utilizar o has utilizado algún sistema de traducción automática (TA) en línea?

☐ Sí

☐ No

¿Cuál(es)? *

☐ Google Translate

☐ SYSTRANet

☐ MateCat

☐ Reverso

☐ WorldLingo

☐ Otros ______________________________

¿Conoces alguna herramienta de posedición?

☐ Sí

☐ No

Annexe IV

Usuario nº:

UNIVERSITÉ DE GENÈVE

CUESTIONARIO PARA PARTICIPANTES

INFORMACIÓN SOBRE LA TAREA

Rellena el cuestionario que se presenta a continuación. Las preguntas con un asterisco (*) al lado del enunciado pueden tener más de una respuesta.

Este cuestionario está enfocado a la tarea de posedición que has realizado. Por favor, lee el cuestionario completo antes de empezar y pregunta cualquier duda que tengas.

SECCIÓN 1: PREGUNTAS GENERALES SOBRE LA TAREA

1. **¿Cuánto tiempo has necesitado para realizar la tarea? Este tiempo puede ser aproximado. Por favor, indica cómo te has organizado para llevarla a cabo, si dividida en varios días o en un solo día, etc., así como cuánto tiempo le has dedicado.**

Días	Horas

2. **¿Ha sido suficiente el plazo para realizar esta tarea?**

___ Sí

___ No

¿Qué plazo hubieses propuesto?

3. **¿Cómo evaluarías la calidad de la traducción automática propuesta?**

Muy buena	Buena	Aceptable	Mala	Muy mala

4. **¿Qué tipo de errores predominan en la traducción automática?**

5. **¿Qué segmentos necesitaban un mayor esfuerzo de posedición?**

SECCIÓN 2: POSEDICIÓN

6. **¿Tienes experiencia como poseditor?**

___ Sí

___ No

7. **¿Qué tipo de experiencia?**

8. **Según tu punto de vista, ¿podría compararse esta experiencia con un encargo real de posedición?**

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

9. **En general, ¿estás de acuerdo con la posedición?**

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

¿Por qué?

10. ¿Crees que es una tarea que perjudica a los traductores?

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

11. ¿Crees que sería necesario introducir algún módulo sobre posedición en el programa de estudios actual?

12. ¿Crees que, en general, el uso de directrices de posedición puede ser útil y/o necesario? ¿Por qué?

13. Las directrices utilizadas, ¿te han resultado útiles?

___ Sí

___ No

14. ¿Eran comprensibles o propondrías algún cambio de formulación?

15. ¿Crees que sería útil añadir ejemplos a las diferentes directrices para facilitar su comprensión y saber qué hay que hacer en cada caso?

___ Sí

___ No

16. Comentarios

¡Gracias por tu tiempo y por completar esta tarea!

Annexe V

Usuario nº:

CUESTIONARIO PARA PARTICIPANTES

INFORMACIÓN SOBRE LA TAREA

Rellena el cuestionario que se presenta a continuación. Las preguntas con un asterisco (*) al lado del enunciado pueden tener más de una respuesta.

Este cuestionario está enfocado a la tarea de posedición has realizado. Por favor, lee el cuestionario completo antes de empezar y pregunta cualquier duda que tengas.

SECCIÓN 1: PREGUNTAS GENERALES SOBRE LA TAREA

1. **¿Cuánto tiempo has necesitado para realizar la tarea? Este tiempo puede ser aproximado. Por favor, indica cómo te has organizado para llevarla a cabo, si dividida en varios días o en un solo día, etc., así como cuánto tiempo le has dedicado.**

Días	Horas

2. **¿Ha sido suficiente el plazo para realizar esta tarea?**

___ Sí

___ No

¿Qué plazo hubieses propuesto?

3. **¿Cómo evaluarías la calidad de la traducción automática propuesta?**

Muy buena	Buena	Aceptable	Mala	Muy mala

4. **¿Qué tipo de errores predominan en la traducción automática?**

5. **¿Qué segmentos necesitaban un mayor esfuerzo de posedición?**

SECCIÓN 2: POSEDICIÓN

6. **¿Tienes experiencia como poseditor?**

___ Sí

___ No

7. **¿Qué tipo de experiencia?**

8. **Según tu punto de vista, ¿podría compararse esta experiencia con un encargo real de posedición?**

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

9. **En general, ¿estás de acuerdo con la posedición?**

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

¿Por qué?

10. ¿Crees que es una tarea que perjudica a los traductores?

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

11. ¿Crees que sería necesario introducir algún módulo sobre posedición en el programa de estudios actual?

12. ¿Crees que, en general, el uso de directrices de posedición puede ser útil y/o necesario? ¿Por qué?

13. ¿Crees que hubieses realizado diferentes cambios de haber tenido una serie de directrices de posedición?

14. Comentarios

¡Gracias por tu tiempo y por completar esta tarea!

Annexe VI

Usuario nº:

UNIVERSITÉ DE GENÈVE

CUESTIONARIO PARA PARTICIPANTES

INFORMACIÓN SOBRE LA TAREA

Rellena el cuestionario que se presenta a continuación. Las preguntas con un asterisco (*) al lado del enunciado pueden tener más de una respuesta.

Este cuestionario está enfocado a la tarea de evaluación de posedición que vas a realizar. Por favor, lee el cuestionario completo antes de empezar y pregunta cualquier duda que tengas.

SECCIÓN 1: PREGUNTAS GENERALES SOBRE LA TAREA

1. **¿Cuánto tiempo has necesitado para realizar la tarea? Este tiempo puede ser aproximado. Por favor, indica cómo te has organizado para llevarla a cabo, si dividida en varios días o en un solo día, etc., así como cuánto tiempo le has dedicado.**

Días	Horas

2. **¿Ha sido suficiente el plazo para realizar esta tarea?**

   ___ Sí

   ___ No

   ¿Qué plazo hubieses propuesto?

3. **¿Qué tipo de errores predominan?**

SECCIÓN 2: EVALUACIÓN DE LA POSEDICIÓN

4. **¿Tienes experiencia como revisor?**

___ Sí

___ No

¿Qué tipo de experiencia?

5. **¿Tienes algún tipo de experiencia como revisor de una posedición?**

___ Sí

___ No

¿Qué tipo de experiencia?

6. **En general, ¿estás de acuerdo con este tipo de práctica, revisor de posedición?**

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

¿Por qué?

7. **¿Crees que es una tarea que perjudica a los traductores?**

___ Completamente de acuerdo

___ De acuerdo

___ En desacuerdo

___ Completamente en desacuerdo

8. **Comentarios**

¡Gracias por tu tiempo y por completar esta tarea!

INSTRUCCIONES PARA LOS PARTICIPANTES

POSEDICIÓN CON MATECAT

Estas instrucciones forman parte del proceso de posedición con la plataforma en línea MateCat.

Por favor, lee detenidamente estas instrucciones antes de comenzar la tarea. Pregunta cualquier duda que se te plantee y/o comenta cualquier problema que se te presente.

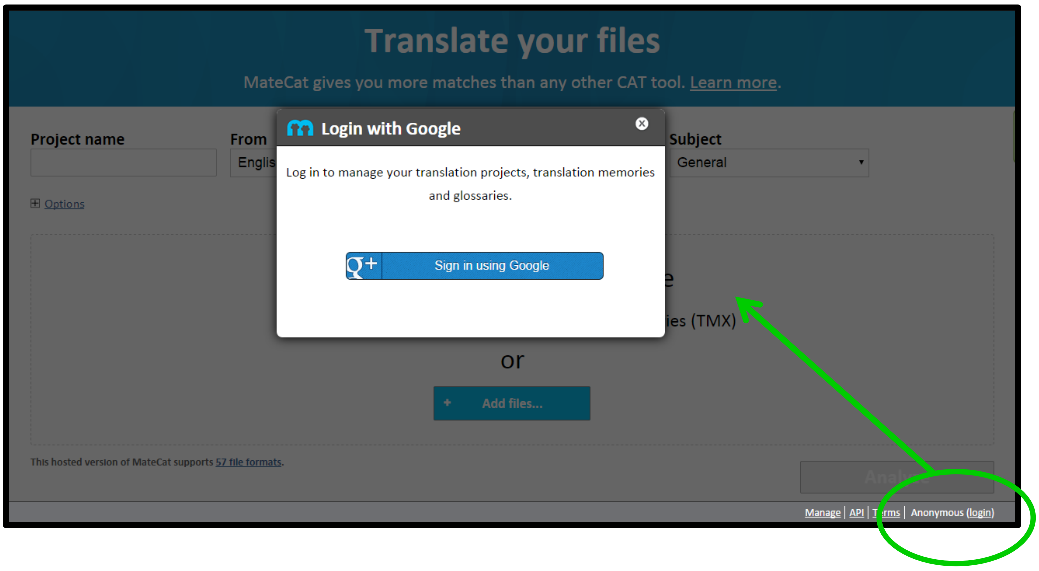
Antes de comenzar, ten en cuenta qué es la posedición. Allen (2003) define la posedición como la edición, modificación y/o corrección de un texto que ha sido pre-traducido por un sistema de traducción automática.

---

1. Abre Google Chrome o Safari y dirígete a www.matecat.com

   La plataforma solo funciona con estos dos navegadores. Si intentamos abrirla con Firefox o Internet Explorer, se nos indica en la página principal que utilicemos Chrome o Safari.

2. Inicia sesión en la barra inferior utilizando la dirección de correo y contraseña que corresponda. Esta información se encuentra en el correo electrónico que se te ha enviado.



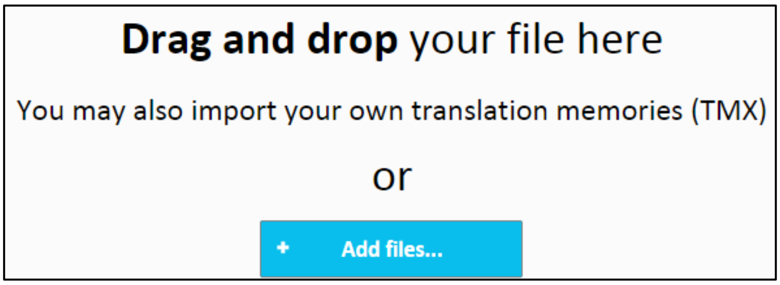
3. Crea un proyecto con los parámetros que se muestran a continuación.

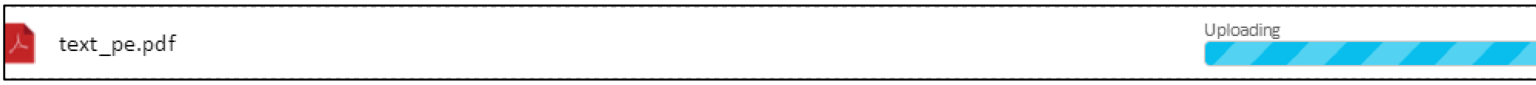
El nombre del proyecto será PE_ seguido del número de usuario que te corresponda, por ejemplo, si eres el usuario 4, el proyecto será "PE_4".

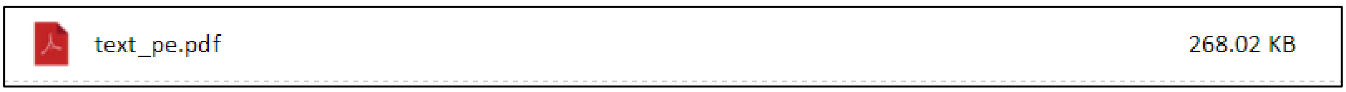
4. Sube el texto a traducir. Encontrarás un archivo PDF con el título "text_pe" en el paquete que se te ha enviado.

Haz clic en 

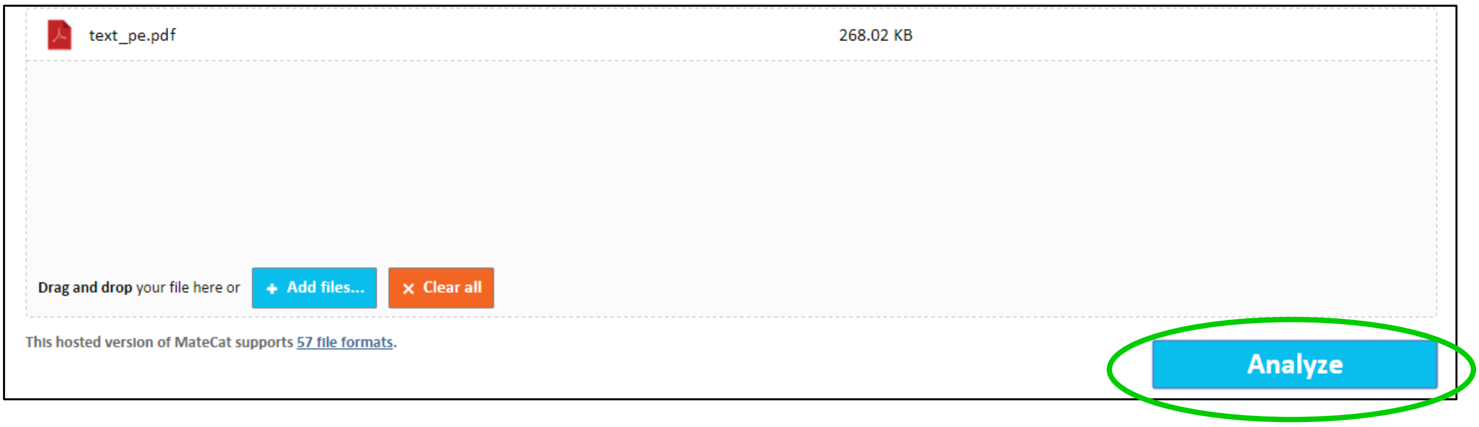
y busca el texto en la ubicación donde hayas guardado los documentos adjuntos.



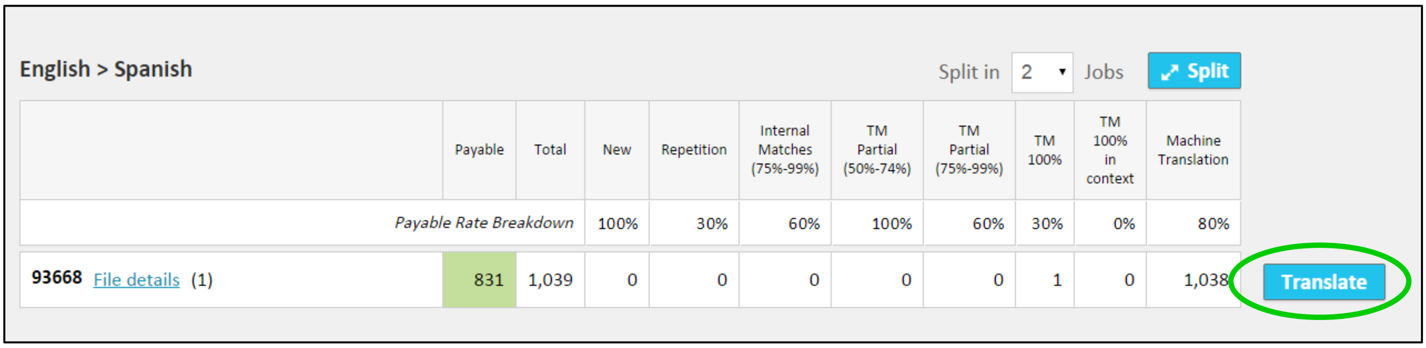




5. Analiza el texto.

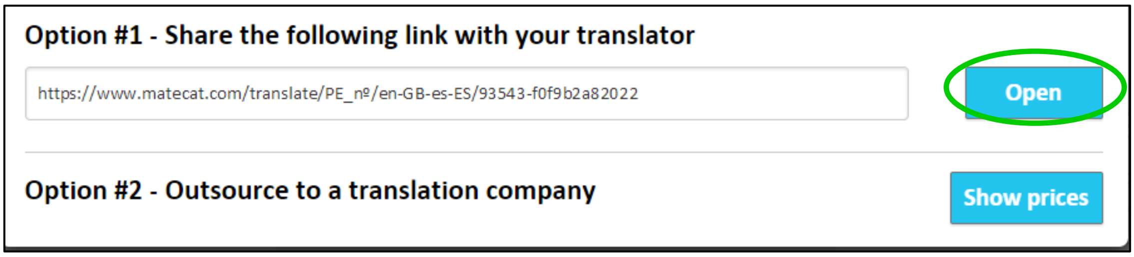


6. Selecciona **Translate** en la siguiente ventana que se abre.

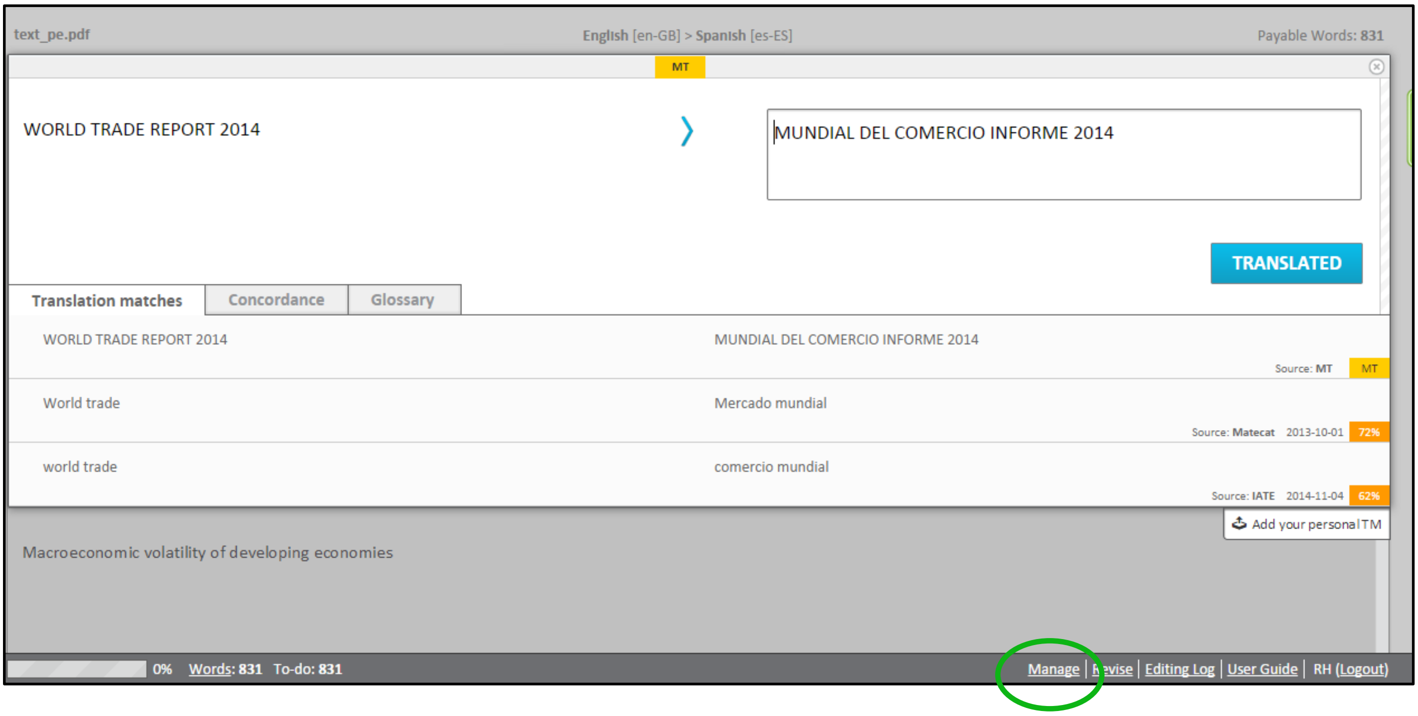


El resultado de tu análisis puede variar al que se muestra en estas instrucciones, el motivo es que si algún participante ha comenzado a trabajar antes que los demás, esto hará que las cifras de resultados TM o MT varíen.

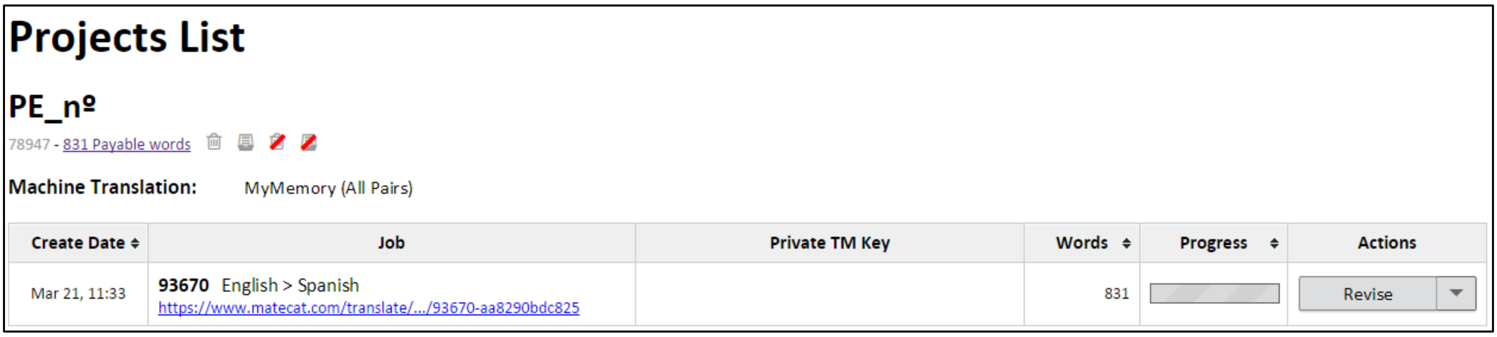
7. Selecciona **Open** en la siguiente ventana. La interfaz de traducción se abrirá a continuación.



8. Antes de empezar a trabajar, dirígete a **Manage**, en la barra inferior, para ver los proyectos activos que tienes.



Verás que aparece este proyecto que acabas de crear titulado “PE” seguido del número de usuario que corresponda:



9. Investiga un poco por la página web para familiarizarte con ella y con la plataforma antes de empezar a trabajar. Puedes encontrar información útil en la **User guide,** el acceso se encuentra en la barra inferior.

En esta página encontrarás información acerca de cómo trabajar con MateCat, así como un vídeo explicativo.

10. Abre el proyecto asignado seleccionando el enlace que aparece en el recuadro.

11. Comienza a poseditar la traducción propuesta siguiendo las directrices de posedición que encontrarás a continuación. También las encontrarás en un documento PDF individual en el paquete de la tarea.

PE Guidelines for publishable quality	
G1	Use as much of the raw MT output as possible, *but*:
G2	Ensure the target reader perfectly understands the content of the target text
G3	Ensure the target text communicates the same meaning and message as the source text
G4	Ensure target-text language is appropriate, but do not restructure sentences solely to improve the flow of the target text
G5	Aim for a grammatically, syntactically and semantically correct target text
G6	Ensure key terminology is correctly translated (considering context and the client’s list of preferred terminology, if available)
G7	Ensure any untranslated terms belong to the client’s list of ‘Do Not Translate’ terms, if available
G8	Edit any offensive, inappropriate or culturally unacceptable content for the target reader
G9	Apply basic rules regarding spelling and punctuation
G10	Ensure the same ST tags are present and in the correct positions in the target text

Puedes hacer los cambios pertinentes en el cuadro de diálogo que aparece a la derecha con la sugerencia de traducción. Una vez hayas realizado los cambios, haz clic en 

para confirmar.

El siguiente segmento para poseditar se abrirá a continuación. Puedes volver a un segmento previo siempre que quieras, simplemente cerrando el segmento abierto en ese momento y haciendo clic en el segmento que quieras abrir.

Dependiendo del segmento y de la base de datos, encontrarás una o varias propuestas de traducción (*translation matches*). Las fuentes variarán entre Matecat, MT y una o varias fuentes externas.
La fuente por defecto es Matecat, por favor verifica que la opción que se muestra es la de traducción automática (**MT**). Es muy importante que **SIEMPRE** sea esta fuente la elegida **en cada segmento**.
Para que sea la sugerencia que aparezca, solo hay que hacer doble clic en la sugerencia **MT**, deberás comprobar la fuente de la traducción en cada segmento. Se puede comprobar que se ha elegido la opción correcta si aparece un pequeño recuadro amarillo con las letras **MT** en la parte superior de la ventana.
La razón es porque si algún otro candidato ya ha realizado cambios, estos serán sugeridos por la plataforma y el resultado de la evaluación no será válido.

Puedes ver un ejemplo en la figura de abajo, la primera sugerencia tiene como fuente a Matecat y 18-03-2015 como fecha. En este caso no ofrece una mejor traducción ya que se debe a pruebas que se realizaron para la configuración de la plataforma, si se hubiesen realizado cambios en la traducción propuesta aparecerían como sugerencias.
Toma tu tiempo y asegúrate de que seleccionas **MT** como fuente.

12. Cuando la posedición se ha terminado y todos los segmentos se han confirmado, puedes cerrar sesión y salir de MateCat.

13. No olvides revisar y rellenar los cuestionarios que se encuentran en el paquete de la tarea.

14. Adjunta los cuestionarios debidamente cumplimentados y manda un correo electrónico para avisar de que has terminado la tarea.

INSTRUCCIONES PARA LOS PARTICIPANTES

POSEDICIÓN CON MATECAT

Estas instrucciones forman parte del proceso de posedición con la plataforma en línea MateCat.

Por favor, lee detenidamente estas instrucciones antes de comenzar la tarea. Pregunta cualquier duda que se te plantee y/o comenta cualquier problema que se te presente.

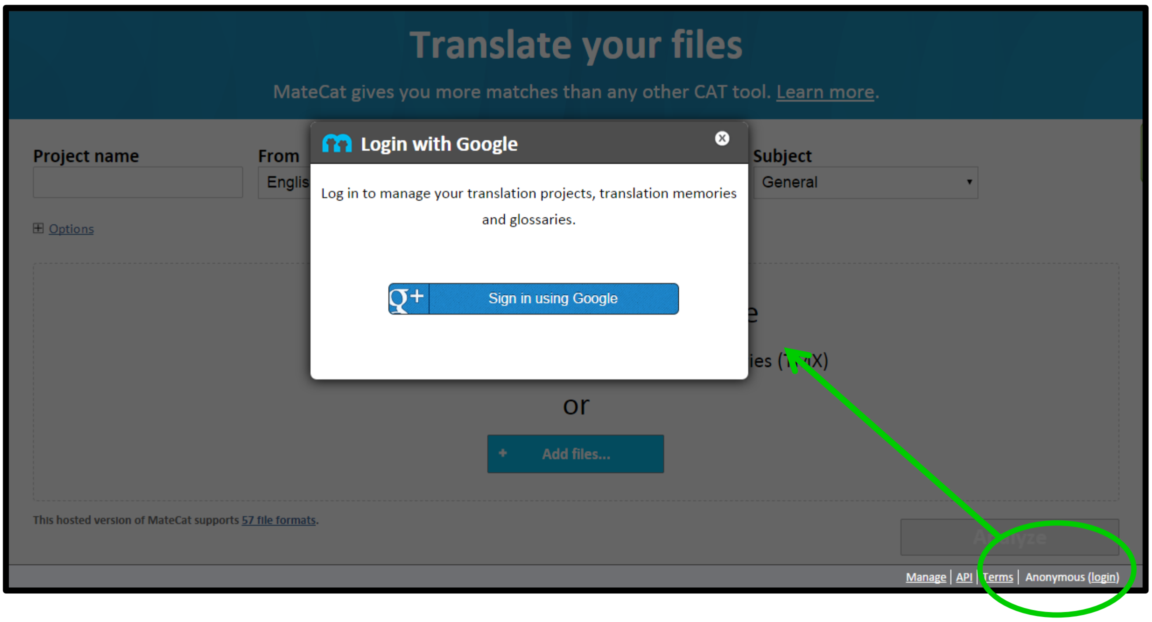
Antes de comenzar, ten en cuenta qué es la posedición. Allen (2003) define la posedición como la edición, modificación y/o corrección de un texto que ha sido pre-traducido por un sistema de traducción automática.

---

1. Abre Google Chrome o Safari y dirígete a www.matecat.com

   La plataforma solo funciona con estos dos navegadores. Si intentamos abrirla con Firefox o Internet Explorer, se nos indica en la página principal que utilicemos Chrome o Safari.

2. Inicia sesión en la barra inferior utilizando la dirección de correo y contraseña que corresponda. Esta información se encuentra en el correo electrónico que se te ha enviado.



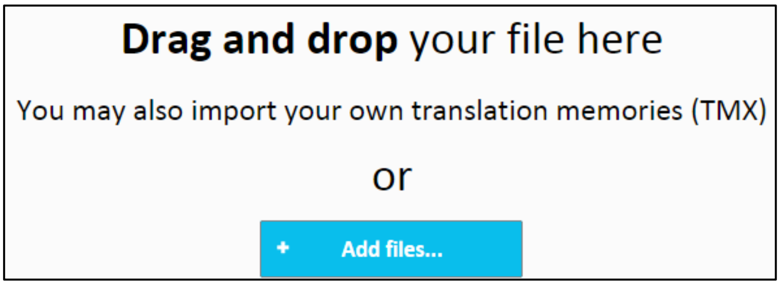
3. Crea un proyecto con los parámetros que se muestran a continuación.

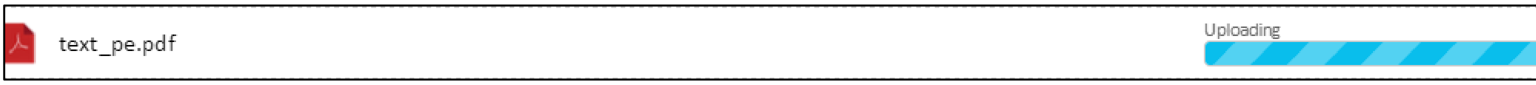
El nombre del proyecto será PE_ seguido del número de usuario que te corresponda, por ejemplo, si eres el usuario 4, el proyecto será "PE_4".

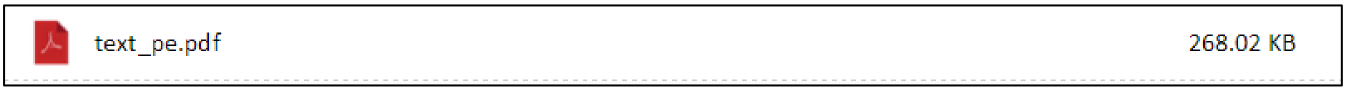
4. Sube el texto a traducir. Encontrarás un archivo PDF con el título "text_pe" en el paquete que se te ha enviado.

Haz clic en 

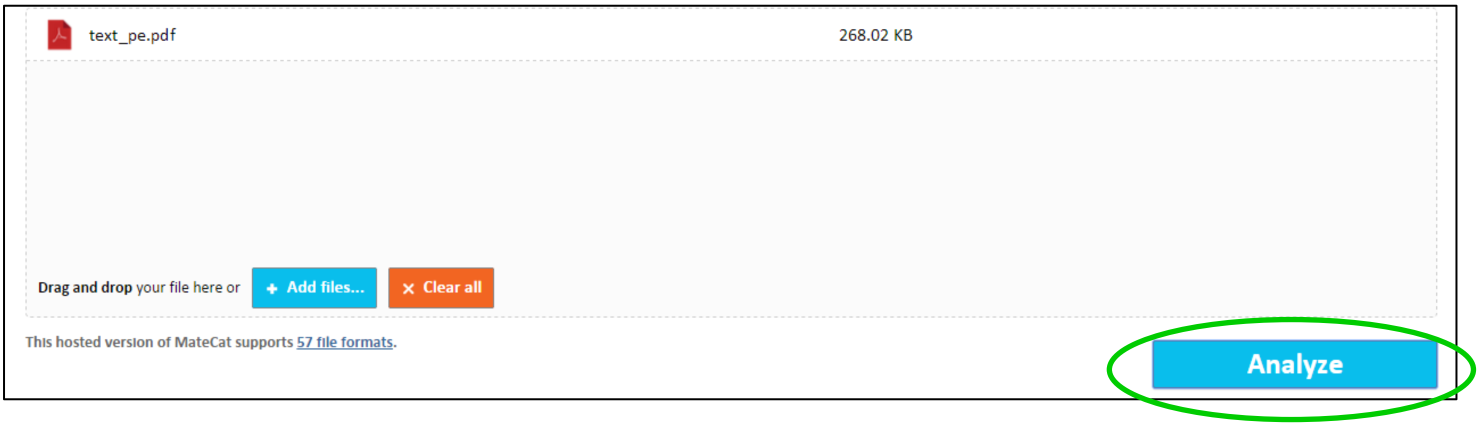
y busca el texto en la ubicación donde hayas guardado los documentos adjuntos.



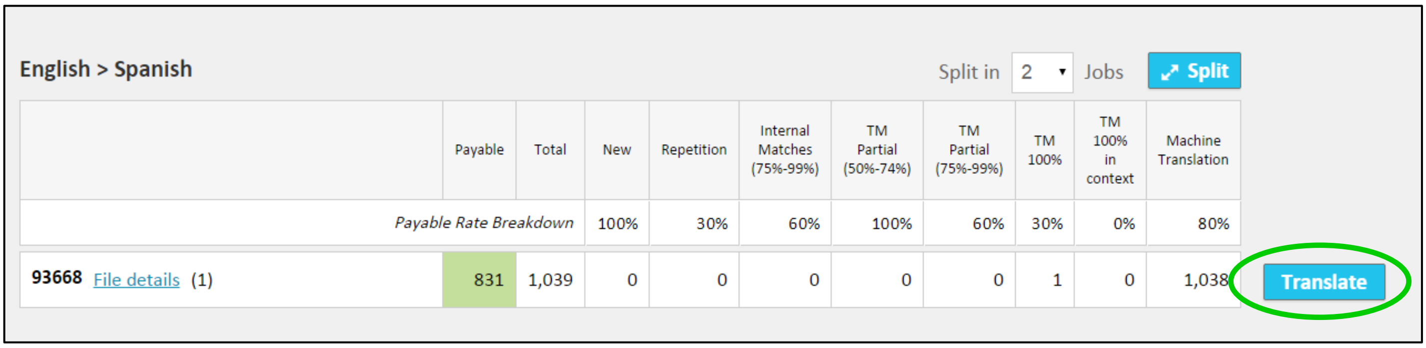




5. Analiza el texto.



6. Selecciona **Translate** en la siguiente ventana que se abre.

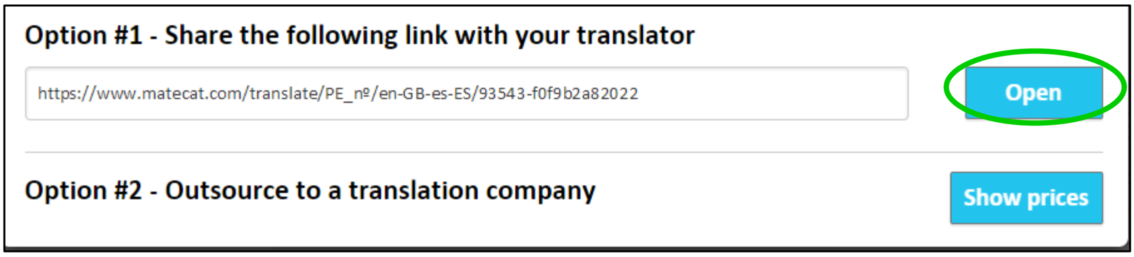


English > Spanish — Split in 2 Jobs — Split

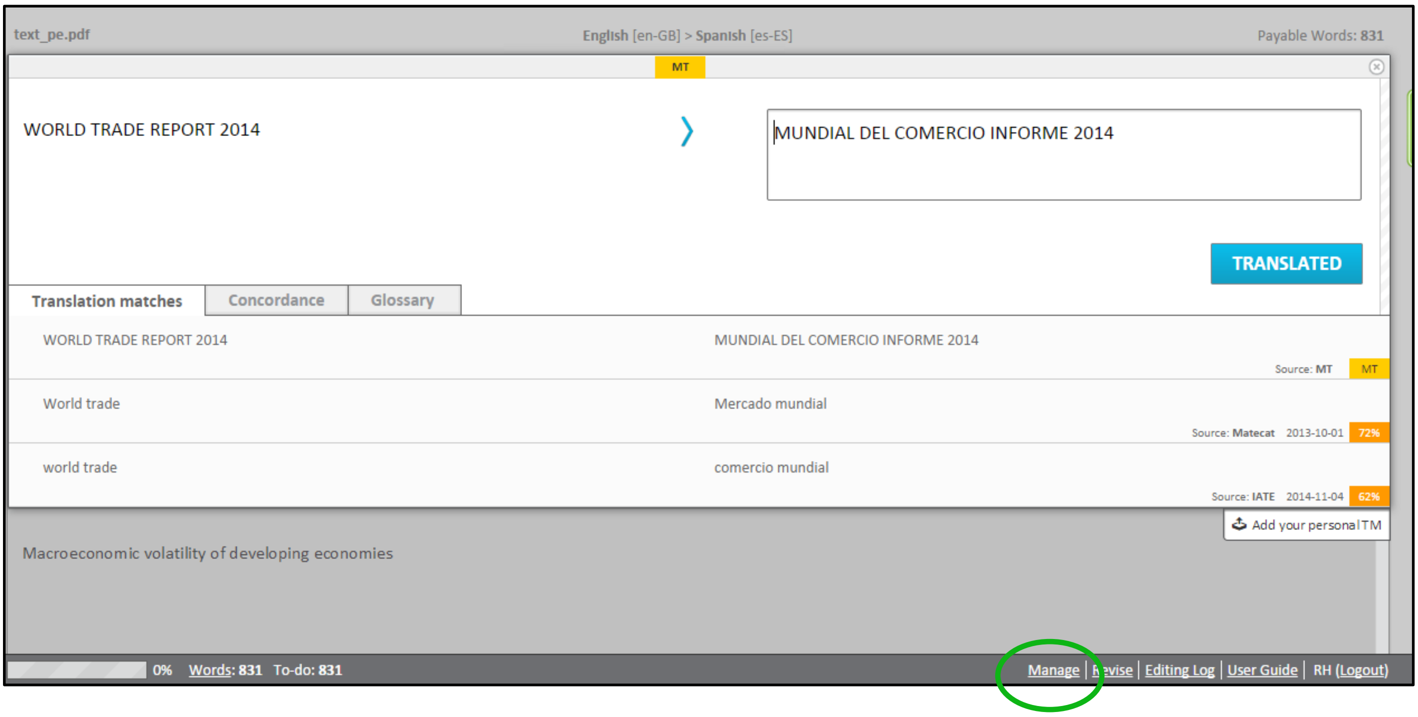
	Payable	Total	New	Repetition	Internal Matches (75%-99%)	TM Partial (50%-74%)	TM Partial (75%-99%)	TM 100%	TM 100% in context	Machine Translation	
Payable Rate Breakdown			100%	30%	60%	100%	60%	30%	0%	80%	
93668 File details (1)	831	1,039	0	0	0	0	0	1	0	1,038	Translate

El resultado de tu análisis puede variar al que se muestra en estas instrucciones, el motivo es que si algún participante ha comenzado a trabajar antes que los demás, esto hará que las cifras de resultados TM o MT varíen.

7. Selecciona **Open** en la siguiente ventana. La interfaz de traducción se abrirá a continuación.



8. Antes de empezar a trabajar, dirígete a **Manage**, en la barra inferior, para ver los proyectos activos que tienes.



Verás que aparece este proyecto que acabas de crear titulado “PE” seguido del número de usuario que corresponda:

Projects List

PE_nº

78947 - 831 Payable words

Machine Translation: MyMemory (All Pairs)

Create Date	Job	Private TM Key	Words	Progress	Actions
Mar 21, 11:33	**93670** English > Spanish https://www.matecat.com/translate/.../93670-aa8290bdc825		831		Revise

9. Investiga un poco por la página web para familiarizarte con ella y con la plataforma antes de empezar a trabajar. Puedes encontrar información útil en la **User guide,** el acceso se encuentra en la barra inferior.

En esta página encontrarás información acerca de cómo trabajar con MateCat, así como un vídeo explicativo.

10. Abre el proyecto asignado seleccionando el enlace que aparece en el recuadro.

11. Comienza a poseditar la traducción propuesta.

Puedes hacer los cambios pertinentes en el cuadro de diálogo que aparece a la derecha con la sugerencia de traducción. Una vez hayas realizado los cambios, haz clic **TRANSLATED** en para confirmar.

El siguiente segmento para poseditar se abrirá a continuación. Puedes volver a un segmento previo siempre que quieras, simplemente cerrando el segmento abierto en ese momento y haciendo clic en el segmento que quieras abrir.

Dependiendo del segmento y de la base de datos, encontrarás una o varias propuestas de traducción (*translation matches*). Las fuentes variarán entre Matecat, MT y una o varias fuentes externas. La fuente por defecto es Matecat, por favor verifica que la opción que se muestra es la de traducción automática (**MT**). Es muy importante que **SIEMPRE** sea esta fuente la elegida **en cada segmento**.

Para que sea la sugerencia que aparezca, solo hay que hacer doble clic en la sugerencia **MT**, deberás comprobar la fuente de la traducción en cada segmento. Se puede comprobar que se ha elegido la opción correcta si aparece un pequeño recuadro amarillo con las letras **MT** en la parte superior de la ventana.

La razón es porque si algún otro candidato ya ha realizado cambios, estos serán sugeridos por la plataforma y el resultado de la evaluación no será válido.

Puedes ver un ejemplo en la figura de abajo, la primera sugerencia tiene como fuente a Matecat y 18-03-2015 como fecha. En este caso no ofrece una mejor traducción ya que se debe a pruebas que se realizaron para la configuración de la plataforma, si se hubiesen realizado cambios en la traducción propuesta aparecerían como sugerencias.
Toma tu tiempo y asegúrate de que seleccionas **MT** como fuente.

12. Cuando la posedición se ha terminado y todos los segmentos se han confirmado, puedes cerrar sesión y salir de MateCat.

13. No olvides revisar y rellenar los cuestionarios que se encuentran en el paquete de la tarea.

14. Adjunta los cuestionarios debidamente cumplimentados y manda un correo electrónico para avisar de que has terminado la tarea.

Annexe IX

UNIVERSITÉ DE GENÈVE

INSTRUCCIONES PARA LOS PARTICIPANTES

EVALUACIÓN HUMANA DE LA POSEDICIÓN

Estas instrucciones forman parte del proceso evaluación de la posedición humana realizada en la plataforma en línea MateCat.

Por favor, lee detenidamente estas instrucciones antes de comenzar con la tarea. Pregunta cualquier duda que se te plantee y/o comenta cualquier problema que se te presente.

Antes de comenzar haremos una pequeña introducción a la evaluación humana del proceso de posedición de un texto traducido por un sistema de traducción automática.

Existen dos formas de realizar esta evaluación: automática y humana. En este caso utilizaremos el segundo tipo.

Siguiendo a Hutchins y Somers, existen varias medidas para evaluar la calidad del texto: precisión (*accuracy*), claridad (*clarity*) y estilo (*style*).

También proponen analizar los errores. Para ello, el poseditor "[...] counts each addition or deletion for a word, each substitution of one Word by another, each instance of the transposition of words in phrases [...]" (Hutchins y Somers 1992, 164).

En nuestro caso mediremos la precisión de la versión poseditada con respecto al texto origen con una escala de 1 a 5, siendo 5 el nivel máximo de precisión. También anotaremos los errores siguiendo la definición del segundo método.

---

1. Abre el documento Excel que se titula "Human PE Evaluation", se encuentra en el paquete de la tarea.

2. Encontrarás un cuadro con las siguientes columnas:

Source Text	Texto origen en inglés
Raw MT Output	Pre-traducción del texto fuente realizada por MateCat, sistema de traducción automática (TA)
PE Version	Versión poseditada de manera humana por uno de los participantes
PE Accuracy 1=not accurate at all, 5= completely accurate	En esta columna se medirá la precisión de versión poseditada con respecto del texto original.
Number of modifications: insertions, deletions, substitutions, shifts, others	En esta columna se anotarán el número de modificaciones realizadas entre el texto pre-traducido con TA y la versión poseditada. Es necesario anotar cuántas veces se ha insertado, eliminado, sustituido o cambiado un elemento. También puedes anotar otro tipo de fenómeno en la columna de “others”
Preferred version: MT output, PE	Elige, según tu opinión, cuál de las dos versiones de la traducción tiene mejor calidad. En caso de que ambas sean iguales, puedes marcar las dos o hacer un pequeño comentario
Comments	Añade cualquier tipo de comentario, por ejemplo, sobre las dificultades o casos poco usuales que hayas encontrado

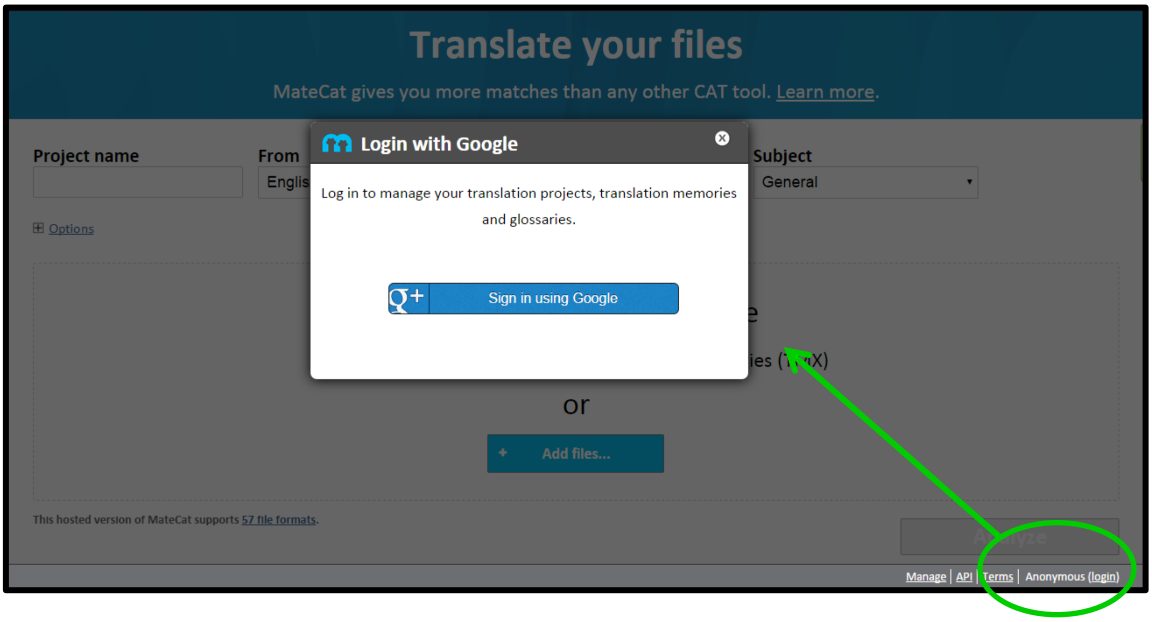
3. Guarda el documento.

4. Comienza con la segunda parte de la evaluación.

5. Abre Google Chrome o Safari y dirígete a www.matecat.com

La plataforma solo funciona con estos dos navegadores. Si intentamos abrirla con Firefox o Internet Explorer, se nos indica en la página principal que utilicemos Chrome o Safari.

6. Inicia sesión en la barra inferior utilizando la dirección de correo y contraseña que corresponda. Esta información se encuentra en el correo electrónico que se te ha enviado.



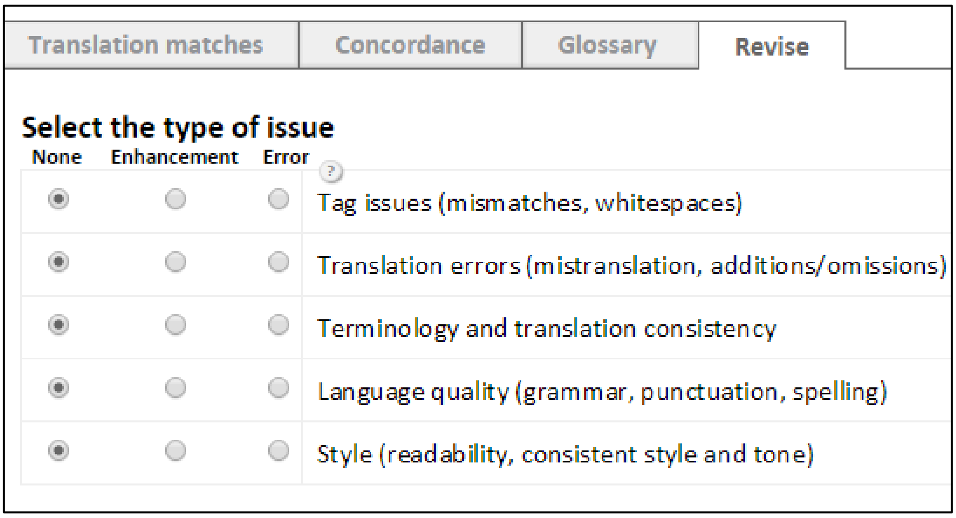
7. Investiga un poco por la página web para familiarizarte con ella y con la plataforma antes de empezar a trabajar. Puedes encontrar información útil en la **User guide,** el acceso se encuentra en la barra inferior.

   En esta página encontrarás información acerca de cómo trabajar con MateCat, así como un vídeo explicativo.

8. Pasa a la pestaña **Revise**, de nuevo en la barra inferior.

9. Señala los tipos de errores que se han encontrado en cada segmento.

No realices cambios, marca solo los errores (si los hay) de los diferentes segmentos.



10. Cuando termines de seleccionar los errores de cada segmento, haz clic en 

11. No olvides revisar y rellenar los cuestionarios que encontrarás en el paquete de la tarea.

12. Adjunta el documento Excel y los cuestionarios debidamente cumplimentados y manda un correo electrónico para avisar de que has terminado la tarea.