Universitat Autònoma de Barcelona

Implementing Machine Translation and Post-Editing to the Translation of Wildlife Documentaries through Voice-over and Off-screen Dubbing

Carla Ortiz Boix

Universitat Autònoma de Barcelona

Departament de Traducció i d'Interpretació i d'Estudis de l'Asia Oriental

Doctorat en Traducció i Estudis Interculturals

Implementing Machine Translation and Post-Editing to the Translation of Wildlife Documentaries through Voice-over and Off-screen Dubbing

A Research on Effort and Quality

PhD dissertation presented by:

Carla Ortiz Boix

Supervised and tutorized by:

Dr. Anna Matamala

2016

A la meva família:

als que hi són,

als que no,

i als que només hi són a mitges.

Acknowledgments

The road to finishing this PhD has not been easy and it would not have been accomplished without the priceless support of many:

First of all, I want to thank my supervisor, Dr. Anna Matamala, for all her hard work. It has not been an easy road and sometimes I would have lost the right path if she had not been there to support, encourage, and challenge me. The PhD would not have come out the way it has without you.

On a professional level, I also want to thank Dr. Pilar Orero for embracing me at the MRA and for all the good moments there. It has been a great pleasure. To the MRA ladies (and men): the ones that I saw leave and the ones that came after me. I learned a great bit from each and every one of you. To the members of TransMedia Catalonia (2014SGR0027) and the Department of Translation, Interpretation and East Asian Studies at UAB with whom I have had the pleasure to share moments. To AGAUR and UAB for awarding me with the FI-DGR2013 scholarship to be able to fulfill this PhD.

I would like to thank specially Manuel Herranz, who welcomed me in Pangeanic Valencia and allowed me to learn from them, and Dr. Sharon O'Brien, for allowing me to do a research stay at DCU, for tutoring me while I was there and for all the great advice. I would also want to thank all the people who voluntarily participated in the experiments; this PhD would not be the same without them all.

On a personal level, I want to first of all thank my parents for all the love and patience they have given me. For being there for me always, no matter what. My life would not be the same without me sister Aurora, who believes in me more than I do and encourages me to be better. I have always been told she looks up to me, but she should realize that she is the most amazing, caring and loving sister and person anyone could ask for. My niece Elna will grow up an extraordinary human being thanks to her and Dani. I have also been surrounded all my life by my grandparents, and I am proud to be their granddaughter.

During this last three years I have not been able to go to Banyoles as much as I would have wanted. I would like to thank my friends for their patience and for dealing with my absences. I owe "tocs" to many that I intend to pay soon. I don't want to forget Cristina and Ainara, my Leipzig family; I will miss you in the USA.

And last, but not least, I want to thank David. He came into my life when I least expected it and he decided to stay. This last year would not have been the same without him next to me. He is my rock, he gives me strength when I need it, and lends me his hand when I give out. He is what I wished for and more. I can't wait to start the next chapter of our lives.

Index

Index of Tables and Figures

TABLES

Chapter 1. Introduction:

Chapter 2. First Article:

Chapter 3. Second Article:

Chapter 4. Third Article:

Chapter 5. Fourth Article:

FIGURES:

Chapter 3. Second Article:

Chapter 4. Third Article:

Chapter 5. Fourth Article:

Acronyms

ALPAC. Automatic Language Processing Advisory Commitee.

AMTA. Association for Machine Translation of the Americas.

ANOVA. Analysis of Variance.

APR. Avarage Pause Ratio.

AV. Audiovisual.

AVT. Audiovisual Translation.

BBC. British Broadcasting Corporation.

BLEU. Bilingual Evaluation Understudy.

CASMACAT. Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation

CAT. Computer Assisted Translation.

EAMT. European Association for Machine Translation.

EBMT. Example-Based Machine Translation.

EMMA. European Multiple MOOC Aggregator.

eTITLE. European multilingual transcription and subtitling services for digital media content.

EU. European Union.

EU-BRIDGE.

H-BLEU. Human-based Bilingual Evaluation Understudy.

H-TER. Human-based Translation Error Rate.

LISA. Localization Industry Standards Association.

LREC. Language Resources and Evaluation Conference.

METEOR. Metric for Evaluation of Translation with Explicit Ordering.

MT. Machine Translation.

MQM. Multidimensional Quality Metrics.

MUSA. Multi-cloud Secure Applications.

NAATI. National Accreditation Authority for Translators and Interpreters.

NIST. National Institute of Standards and Technology.

OD. Off-screen Dubbing.

PE. Post-editing.

PE MT. Post-edited Machine Translation.

PET. Post-Editing Tool.

PWR. Pause to Word Ratio.

QA. Quality Assessment

SL. Source Language.

SUMAT. Subtitling by Machine Translation.

TAUS. Translation Automation User Society.

TER. Translation Error Rate.

TC. Time Code.

TSP. Translation Service Providers.

TransLectures.

VO. Voice-over.

WER. Word Error Rate.

Chapter 1. Introduction

1. Introduction

"The overall picture is this. The more Machine Translation there is, the more translation will happen, the more people will expect to be able to communicate with other folk, and the more they will realize that although machines can clear the ground, the actual translation has to be done by somebody because language is human behavior. It's machine simulated, but they're not doing anything like what a human translator is doing."

(David Bellos in Erickson, 2012)

"[t]he media, new technologies and human and automatic translation services can bring the increasing variety of languages and cultures in the EU closer to citizens and provide the means to cross language barriers"

(European Comission, 2008:12)

Accessibility can be considered a universal right that does not want to exclude anyone because of lack of linguistic knowledge or disability in our multilingual and diverse society. Traditional media, as well as new and digital medias, deliver information and entertainment through various channels and in different languages. In order to make these products accessible for everyone, audiovisual transfer modes such as subtitling for the deaf and hard-of-hearing, audio description or sign language interpreting can be used, so that sensorial accessibility can be achieved; and dubbing, subtitling or voice-over, among other transfer modes, can be implemented to overcome linguistic barriers (Matamala and Orero, 2007). All these audiovisual transfer modes have mostly been based on human translations, but as it is stated by the European Commission (2008), not only human translation but also machine translation provide the means to cross language barriers, and thus, increase linguistic accessibility among speakers of different languages.

The use of Machine Translation (MT) among translation companies has grown exponentially during the last decade and will continue to grow at the same level in the near future (Van der Meer and Ruopp, 2014). Many translation professionals, however, are biased against the use of MT (Guerberof Arenas, 2013; Gaspari, 2001) because they are concerned their rates will decrease in order to be competitive against the price asked for a translation produced by a machine, and because they think that MT engines may endanger their profession. However, as David Bellos, director of Princeton's Program in Translation and Intercultural Communication, stated during a discussion on what he believes is lost and found in MT (Erickson, 2012), the growing use of MT will help increase the demand of translations, and hence, what Matamala and Orero (2007) have termed linguistic accessibility. At the same time, he seeks to make people realize that MT can only clear the ground and human translators will always be needed, either as traditional translators or as post-editors, for language is inherent in the human behavior.

The growing use of MT, though, has not impacted all types of translation equally (O'Hagan, 2007). In the case of Audiovisual Translation (AVT), researchers have so far only studied the possibility of implementing MT and post-edited machine translation (PE MT) in AVT's written modality, subtitling (e.g. Georgakopoulou and Bywood, 2014), thanks to EU financed projects like SUMAT (2011-2014), MUSA (2002-2004), eTITLE (2004-2006), and EU-BRIDGE (2012-2014), TransLectures (2011-2014), or EMMA (2014-2016). This implementation has achieved promising results that could translate into the use of PE MT by translation companies in the future. The possible implementation of MT and PE MT is just starting to be researched in oral AVT modalities such as audio description (Matamala, 2015; Fernández-Torné, forthcoming).

The research carried out in this PhD intends to fill this gap and study the possible implementation of MT and PE MT into voice-over (VO) and off-screen dubbing (OD), which are two tightly related oral AVT modalities used to make non-fiction audiovisual products linguistically accessible in countries such as Spain. In order to do so, the inclusion of MT into the translation of documentaries of a specific domain –wildlife– has been analysed, studying not only what challenges need to be addressed, but also by proving whether the inclusion of PE MT would be appropriate and valuable in terms of effort and quality.

Wildlife documentary films have been chosen as an exponent of non-fiction audiovisual products. Particularly, the challenges that need to be bested to include MT into the translation of wildlife documentary films and propose solutions will be studied. Furthermore, the effort required to post-edit machine translated wildlife documentaries as opposed to translation from scratch will be analyzed through an experimental study. Finally, by means of another experimental study and a user reception study, the quality of the post-edited documents will be assessed and compared to the quality of human translations.

As research in post-editing (PE), machine and audiovisual translation is interdisciplinary, this dissertation presents a mixed method study. It combines qualitative and quantitative approaches, as well as different sets of data, while approaching the possibility of including PE MT into a so far unexplored field: documentary film translation through voice-over and off-screen dubbing. It has been developed within the wide framework of Translation Studies, particularly in Audiovisual Translation Studies, but it also relates to the field of language technology, more specifically to machine translation and post-editing, and to the field of translation quality evaluation. Hence, concepts and methods from translation process research and translation quality analysis have been imported into the field of Audiovisual Translation to establish a sound theoretical and methodological framework, as described below.

This research has been done thanks to ALST project (Technologies for Linguistic and Sensorial Accessibility), which intends to apply translation and speech technologies to two AVT transfer modes, namely audio description (as an instance of sensorial accessibility) and voice-over/off-screen dubbing (as an instance of linguistic accessibility), in order to determine whether such technologies can help increase both types of accessibility in Spain. After the promising results reported in an MA thesis on the inclusion of both translation and speech technologies into the audio description process (Ortiz-Boix, 2012), and the first studies on speech technologies for audio description in Catalan (Fernández-Torné and Matamala, 2014), there was still a need for a research working on the inclusion of such technologies into other oral transfer modes, namely voice-over and off-screen dubbing. This PhD, however, only focuses on the inclusion of MT, as the results achieved in the first experiment, which calculated the post-editing effort in comparison to the effort of translation, caused the need for deeper research in that topic.

This PhD is registered in the Translation and Intercultural Studies PhD Program (Doctorat en Traducció i Estudis Interculturals) of the Universitat Autònoma de Barcelona's Department of Translation and Interpretation and East Asian Studies (Departament de Traducció, Interpretació i Estudis de l'Àsia Oriental). It was awarded the three-year FI-DGR2013 scholarship by the Catalan Government's Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) and it has been funded by the Spanish Ministry of Ciencia y Competitividad's ALST Project (FFI-2012-31024), led by Dr. Anna Matamala, within the TransMedia Catalonia Research Group (2014SGR27).

1.1. Research Questions, Objectives and Hypothesis

After a preliminary study presented on my MA thesis, in which it was determined that MT might be implemented in audio description in the Catalan>Spanish language pair if the output of the MT engine was post-edited, the need to research whether PE MT might be included in other AVT oral transfer modes, such as VO, arose. Hence, this dissertation was conceived to answer an initial research question:

- ♦ Is it worth including MT and PE MT into the process of translating audiovisual (AV) content through voice-over and off-screen dubbing?

However, as VO is used to translate several AVT products and a single PhD can not encompass them all, I narrowed down the subject of study and selected a product that is broadly translated by means of VO: documentaries. Furthermore, in order to enclose even more the research, a type of documentary film which could be representative of all the other types was chosen: wildlife. Hence the initial research question evolved and led me to the main objective of this PhD:

- ♦ Research whether machine translation might be successfully included, effort and quality wise, into the process of translating documentaries of a certain subdomain (wildlife) through voice-over and off-screen dubbing.

Two hypotheses arose from this objective:

- ♦ The inclusion of MT into the process of translating wildlife documentaries through VO and OD will optimize the process in terms of effort.
- ♦ The inclusion of MT into the process of translating wildlife documentaries through VO and OD will not have a significant impact on the quality of the translated product.

However, in order to fulfill the main objective of this PhD, other research questions needed to be addressed: e.g. what are the specific characteristics of wildlife documentaries and how are they going to impact the inclusion of MT into the process? How do the characteristics of VO translation affect the normal process of machine translating and post-editing? How much effort is required to post-edit wildlife documentaries? Is this effort lower than translating from scratch? Is the quality of the post-edited wildlife documentaries similar to the quality of the translations? Do the users notice any differences between post-edited and translated documentaries?

Therefore, three secondary objectives were established:

- Determine what characteristics of documentary translation through VO and OD would be challenging when including MT into the process and propose solutions to overcome such challenges.

- Compare the amount of effort required to post-edit machine translated documentaries to be voiced-over and off-screen dubbed to the effort of translating them from scratch.

- Assess the quality of post-edited documentaries to be voiced-over and off-screen dubbed as compared to the quality of translated documentaries.

Two hypotheses arose from the last two secondary objectives:

- The effort of post-editing machine translated documentaries to be voiced-over and off-screed dubbed will be significantly lower than the effort of translating them from scratch.

- The inclusion of translation technologies into the process of translating wildlife documentaries through VO and OD will not affect the quality of the translated product and, hence, the quality of post-edited documentaries is comparable to the quality of documentaries translated from scratch.

This dissertation is presented as a compilation of articles (see structure on subsection 1.4.), and every article presented addresses a secondary objective and helps to validate the main hypothesis of this PhD. In the following table, the correlation between objectives and articles can be observed:

Art. 1	Art. 2	Art. 3	Art. 4	Objective
√				Determine what characteristics of documentary translation through VO and OD would be challenging when including MT into the process and propose solutions to overcome such challenges.
	√			Compare the amount of effort required to post-edit machine translated documentaries to be voiced-over and off-screen dubbed to the effort of translating them from scratch.
		√	√	Assess the quality of post-edited documentaries to be voiced-over and off-screen dubbed as compared to the quality of translated documentaries.
√	√	√	√	Research whether machine translation might be successfully included, effort and quality wise, into the process of translating documentaries of a certain subdomain (wildlife) through voice-over and off-screen dubbing.

Table 1. *Correlation Articles-Objectives*

1.2. Theoretical Framework and Research Background

Because of the multidisciplinary nature of this PhD, this section is divided in three subsections, one for each field of study comprised in it: audiovisual translation, language technologies (MT and PE), and quality assessment. The first subsection has its focus on the modality and genres under analysis in this PhD: voice-over and off-screen dubbing in wildlife documentaries. This is why research on voice-over in fiction genres (Sepielak and Matamala, 2014; Sepielak, 2014) is not approached. In the second subsection, the state of the art of machine translation, giving special emphasis to post-editing and post-editing effort, is described. The third and last subsection presents key concepts in quality assessment, both regarding translation and post-editing.

Some of the information presented in this section in a cohesive and structured way is also included separately in the articles. The result is that at some points the dissertation may seem repetitive, a situation that has been unavoidable due to its presentation by compendium of articles.

1.2.1. *Translation of Documentaries through Voice-Over and Off-Screen Dubbing*

Voice-over, often in combination with off-screen dubbing, is the AVT transfer mode applied in many West European, and Central and South American countries to factual genres –e.g. news, documentaries and documentary series, talk and reality shows, or political debates–. Such genres often intend to portray reality and support the arguments presented as true and trustworthy by relying on visual and verbal evidences (Franco et al., 2010: 24-25). While the visual evidence is displayed through images of events, people, documents, etc.; the verbal evidence is presented via interviews with either experts on the subject matter of the program, also called the voice(s) of expertise (Franco et al., 2010), or testimonies of the subject matter and/or ordinary people with some sort of experience on the subject matter, also known as the voice(s) of experience (Franco et al., 2010). Because of the defining features of VO, which "contribute to the appeals of reality, truth and authenticity that factual programs count on in order to prove their arguments as right or believable" (Franco et al., 2010: 25), the voices of experience and expertise are usually the material to be translated through VO.

The term VO as a mode of transfer in the field of AVT appears for the first time in Fawcett (1983; in Franco et al., 2010), where it is described as a form of dubbing. However, it is not until the late 1990's that VO was considered another AVT transfer mode *per se*. Actually, it is not until the early 2000´s when the first systematic study on VO was published. In her research, Franco (2000) exposes the small amount of documented research work pertaining to either VO or the factual genre. Accordingly, only 2.3% of the entries (29 out of 1241) within the second edition of *Language Transfer and Audiovisual Communication Bibliography* (Gambier, 1997) were devoted to factual programs and in only another 0.9% of the entries (11 out of 1241) the term “voice-over” appeared explicitly. Hence, only 3.2% of the entries (40 out of 1241) referred explicitly to the factual genre and VO, including multiple entries of Pönniö’s (1995) and Luyken et al. (1991), as their studies contained more than one topic.

Orero (2006) presented an update of the data in Franco (2000) at the MuTra Conference in Copenhagen. In Orero (2006) the used data was extracted from 4 online bibliographies: *Translation Studies Bibliography* (John Benjamins) <http://www.benjamins.com/online/tsb>[1], *Translation Studies Abstracts* and *Bibliography of Translation Studies* (St. Jerome) <http://www.stjerome.co.uk/tsaonline/index.php>[2], and *Bibliografia de Traducció i d’Interpretació*, BITRA, (Javier Aixelà, Universitat d’Alicante) <http://aplicacionesua.cpd.ua.es/tra_int/usu/buscar.asp?idioma=va>[1]. Although the number of research works on VO increased, only 8 more references were documented, a small proportion when compared to the other transfer modes –subtitling and dubbing–, which translated to an increase of a 0.5% of the total number of entries.

Four years later, Franco et al. (2010) present updated data again. The results show 72 research studies that deal with VO divided between works published before and after Franco (2000). Their findings indicate that the amount of studies related to VO increased substantially and proved that the number of works focused only on VO augmented significantly within a 6-year time framework –23 out of 33 works with VO as a focus were published after 2000–. After Franco et al. (2010), more research studies focused on VO have been published. Despite the amount of works published that deal with VO increasing in the last two decades, it is yet very little compared to the studies published on subtitling or dubbing.

Most studies on voice-over have taken a descriptive approach, which has allowed identifying the main characteristics of VO. Among these descriptive studies, Franco et al. (2010) has been adopted as the framework of this

[1] URL last retrieved in January 2016.

[2] URL last retrieved in February 2010.

dissertation, as it contains a global and detail overview on every aspect of this transfer mode. Despite the increasing number of reception studies in AVT, reception studies that focus on VO and documentaries are mostly nonexistent. Hence, this PhD contributes to this line of investigation by presenting a reception study to assess the quality of a post-edited wildlife documentary, as compared to the quality of a translated wildlife documentary, according to the end-user perspective (see Section 5).

One recurrent aspect found in the literature is the discussion about the authenticity transmitted by voice-over. In their works, authors such as Luyken et al. (1991), Gambier (1996), Scannell (1996), and Carroll (2004) emphasize the faithfulness and literality of VO translation to the original speech. Luyken et al. (1991), Espasa (2004), Garcarz (2006), and Franco et al. (2010) also underline the sense of authenticity of VO translation, as compared to dubbing. However, authors such as Franco (2000), Krasovska (2004), Kaufmann (2004), Franco et al. (2010) and Darwish and Orero (2014) state that, despite the sense of authenticity of VO translation and its relationship to the factual genre, the implied faithfulness, literalness, and actual meaning of the translations do not always happen.

Synchrony and the way voice-over is delivered have also been key aspects in voice-over research (e.g. Kaufmann, 1995; Moreau, 1998; Kovacic, 1998; Orero, 2006; or Zinik, 2006). The main characteristic of VO delivery is that the translating voice is recorded on top of the original voice, which can still be heard. Kauffmann (1995, 2004) adds the fact that the translating voice is mainly used in sequences with dialogues (interviewers/interviewees) or monologues (talking heads). She also determines as important to leave a few seconds at the beginning and at the end of each dialogue entry for the original voice to be heard on its own, also named by Luyken et al. (1991) and Gambier (1996) as “almost synchrony”. However, not all the authors agree with this concept. It is claimed that the original voice is not always left to be heard for some seconds before the voiced-over voice starts and after it finishes, and if it does, the amount of time left to sound varies (Matamala and Sepielak, 2014). The delivery of voiced-over translations does not need to be completely synchronized with the original version but it needs to be synchronized enough, as the original voice can still be heard, although its volume is lowered, which might allow the audience to spot inadequacies. As Orero (2006) exposes, VO synchrony contains the so-called voice-over isochrony, which is different from dubbing synchrony insofar there is no lip synchronization in VO delivery. VO synchrony also contains kinetic and action synchronies, synchrony types that can also be found in dubbing synchrony and are especially important in VO and OD (Franco et al., 2010: 81-82). Kinetic synchrony is described as the synchrony between the body movements appearing in the visuals with the audible translation; in other words, when an on-screen speaker is pointing at something, the translation has to make reference to it in that exact moment instead of doing it before or after it. Action synchrony is the synchrony between the images appearing on screen and the audible translation, which might limit syntactic inversion and impose the order of the speech.

Other linguistic aspects have been analysed regarding voice-over. The adequacy of including markers of oral discourse in the voiced-over translation was a matter of discussion from the early stages. According to some authors, e.g. Luyken et al. (1991), Matamala (2009a, 2009b), or Franco et al. (2010), such markers, as well as mimetic reproduction of accents, are better being eliminated in the VO translation in order to help synchrony, content relevance and credibility, and keep the voice talent less visible.

In non-fictional products, VO is usually combined with off-screen dubbing, sometimes also referred as commentary or narration (Pönniö, 1995), with slightly diverging definitions. OD, which is the translation of off-screen narrations in which the original soundtrack is deleted and substituted by the target language version, shares some characteristics with dubbing, as only the translating voice is delivered. However, it also shares characteristics with VO, especially the genre in which it is applied, as no lip synchronization is needed and the constraints of action synchrony need to be met (Franco et al., 2010). Research on off-screen dubbing has often been combined with research on voice-over.

All in all, research on VO and OD has mostly been descriptive. Departing from the descriptive studies presented here, this PhD also contributes to this field: it expands the set framework, and it studies (for the first time) the possibility to include MT and PE MT into the VO and OD translation process. Furthermore, it goes a step further by doing both experimental and research studies.

Although VO and OD are used to translate all types of factual programs, this PhD only focuses on the translation of a specific type: documentaries. Because of the versatility of documentaries and their hybrid protean nature (Espasa, 2004), only a type of documentary, which can be considered an exponent of all the others, has been selected: wildlife. Espasa (2004) explores the challenges to translate and research in this field from an academic point of view. Espasa, as well as other authors (e.g. Franco, 2001a, 2001b; Matamala, 2009a, 2009b, 2010; or García Luque, 2011), emphasize aspects such as the level of specialization of the vocabulary and terminology used by some of the speakers, as compared to the non-specialization of the vocabulary used by others. According to her, documentaries should be translated similarly to technical and scientific texts and having into account the different degrees of specialization one can find. Other aspects that are addressed by the authors are the variety of registers used within a same documentary or the synchrony, which are considered problematic. Espasa (2004) also highlights the interplay between image and sound and between verbal and non-verbal elements found in a documentary. And remarks the antithesis between documentaries and texts if assuming that documentaries are audiovisual by nature and texts are surmised to be written.

Matamala (2009a), on the other hand, carries out a descriptive study based on her professional experience translating wildlife documentary films and concludes that the some important features of the translation of documentaries are:

(1) *The working conditions of the translators,* as they are to work against the clock, which affects the documentation process. There may be a lack of postproduction scripts or, if they are available, they are of poor quality, having errors, inaccuracies and linguistic inconsistencies (also in Franco et al., 2010).

(2) *the speakers and translation modes*; meaning that, as there are different types of speakers and several techniques are used when translating a documentary, several transfer modes can be found within a same documentary.

(3) *terminology,* as some documentaries could be considered semi-specialized texts, which means that translators need to do research and terminological searches in specialized areas (also in Matamala, 2010).

These characteristics, along with the characteristics of VO and OD presented earlier on this subsection, set the basis of the first article comprised in this PhD, which contains more information on the features presented above.

1.2.2. *Machine Translation and Post-Editing*

Research on MT started over 50 years ago with a clear goal: create a full automatic MT engine that could produce high quality translation, aka Full-Automatic High Quality Translation (Bar-Hillel, 1960). The publication of the ALPAC report in 1966, which claimed that the quality of the MT engines built so far was low and gave no perspectives of improvement, caused the termination of funding devoted to the research on MT. In the 1980's an alternative approach was taken: computers were to be used as tools for the translators instead of being an alternative to them. Since then, great improvements have been made and the implementation of MT into the translation process has been proven particularly successful in domain specific texts, which guarantees translation of better quality for the post-editors to work with (e.g. Isabelle et al., 2007; Offersgaard et al., 2008; Ceasusu et al., 2011; Läubli et al., 2013; Bouillon and Spechbach, 2016), and in general texts, where MT is used for gisting purposes and interpersonal communication (Ray, 2004: 8-9).

Despite the improvements made, few studies have researched the possible application of MT into the process of translating audiovisual products. So far, it has only been researched for two transfer modes: subtitling (e.g. Melero, 2006; Armstrong et al., 2006; Volk et al., 2010; Sousa et al., 2011; Bywood, 2013; Etchegoyhen et al., 2014; Volk, 2014) and, to a lesser extent, audio description (Fernández-Torné et al., 2013; Matamala, 2015; Fernández-Torné, forthcoming). Several studies, however, have researched the possibility of using subtitling corpora in order to create or improve in-domain MT engines, and for research purposes (e.g. Hardmeier and Volk, 2009; Volk et al., 2010; Daumé III and Jagadeesh, 2011; Cettolo, et al., 2012; Ziemski et al., 2016; Lison and Tiedermann, 2016).

Since the rebirth of the research on MT, PE has been in use in organizations like the European Union and the Pan-American Health Organization (García, 2011) and has achieved more and more recognition and interest both in business contexts (e.g. Plitt and Masselot, 2010; Zhechev, 2014; Silva, 2014; Van der Meer and Ruopp, 2014) and in academia, as there are workshops (e.g. Annual Workshop on PE Practice and Technology since 2012) and tracks within large conferences (e.g. MT Summit; LREC, EAMT and AMTA Conferences, etc.) devoted to it.

In business contexts, research has been dedicated to determining whether the use of MT and PE MT helps increase productivity, understood as "translating more pages in a shorter time, with lower costs" (Koponen 2016: 11). Even though it has been proven that PE MT can increase significantly –by 2,000 to 3,500 words per day (Robert, 2013) or by 74% words per hour (Plitt and Masselot, 2010)–, such increment seems to depend on the language pairs and the projects under consideration (e.g. Guerberof Arenas, 2010; Plitt and Masselot, 2010; Zhechev, 2014), and on the experience of the post-editor, the usage of in-domain MT systems or the pre-editing of the source text (García, 2011).

PE could be defined as the human correction of an automatically translated text –raw machine translation– until the translation is acceptable according to a set of specifications. It could be classified in different groups depending on basically four aspects:

(1) Use of a source text:

1. Bilingual post-editing or post-editing: a post-editor corrects the machine translation comparing it to the source text.
2. Monolingual post-editing: a post-editor corrects the machine translated text without having access to the source text.

(2) Person performing the task:

1. Professional post-editing: a professional translator or post-editor corrects the machine translated text.
2. Non-professional crowd-sourced post-editing: users of certain forums or social media correct machine translated user generated content for information purposes.

(3) Purpose of the translation:

1. Informative post-editing: a post-editor corrects a machine translated text for information or in-company purposes, which do not require a high-quality translation.
2. Ready-to-publish post-editing: a post-editor corrects a machine translation for publishing purposes, which requires high quality.

(4) Level of intervention required:

1. Light or rapid post-editing: a post-editor checks the translation to guarantee it contains no mistranslations or offensive content.

2. Medium or minimal post-editing: a post-editor corrects the machine translation by ensuring its meaning and readability.

3. Full post-editing: a post-editor corrects the machine translation text guarantying it contains no grammar, fluency, terminology, style or voice problems.

The selection of one type of post-editing or another inevitably impacts both the quality of the final product and the effort required to carry out the post-editing task. This dissertation intends to analyze the effort required to post-edit a wildlife documentary that is ready to send to the dubbing study –aka ready to publish–, as well as to assess its quality. Hence, in order to accomplish the main objective of this PhD, the participants of the experiment were asked to perform a bilingual, professional, ready-to-publish, and full post-editing.

Effort has been a key research issue in the field of PE since the beginning of the 2000's, mainly thanks to studies such as Krings (2001), Martínez (2003), O'Brien (2004, 2005 and 2006), Englund Dimitrova (2005), Carl et al. (2011), Tatsumi et al. (2012), Lacruz et al. (2014), or Almeida and O'Brien (2010), and Guerberof Arenas (2009), who compared PE effort and translation effort in order to determine which option is more feasible in terms of productivity. Results show that, in the majority of the cases, PE requires less effort than translation from scratch. Regarding PE effort, Specia (2011) researched the possibility to predict PE effort automatically by comparing sentences that were predicted to be of good or average quality. According to the results, sentences predicted as good quality are faster to post-edit than the others.

Krings (2001: 178) set the standard for the majority of the other works on this topic, including this PhD, by presenting a way to calculate PE effort. He divides PE effort into three categories: temporal effort –time required to post-edit a machine translated document–, technical effort –number of keystrokes, mouse movements, and mouse clicks needed to post-edited a machine translated document–, and cognitive effort –cognitive processes required to correct the errors of a machine translated document–. As Krings (2001) himself claims, temporal effort is the most visible aspect of PE. It can be seen in e.g. Allen (2001), Martínez (2003) or Tatsumi and Roturier (2010), that not only temporal effort but also technical effort can be directly observed using keylogging software. This dissertation also contributes to this field by partially replicating studies on PE effort, as compared to translation, using keylogging software, more specifically Inputlog (Leijten and Van Waes, 2013).

However, determining cognitive effort cannot be observed directly, as can be proven by the many different methods used to determine it (e.g. Krings, 2001; O'Brien, 2004; O'Brien, 2006; Shreve et al., 2011; Lacruz et al., 2014). For the study on PE effort presented on this PhD, Lacruz et al.'s (2014) pause-to-word-ratio (PWR) method to determine cognitive effort has been implemented. More information on PWR and how it has been applied in its experimental study can be found in the second article presented in this PhD (see chapter 3).

Despite the increasing number of studies on post-editing and post-editing effort, only a few have been applied to AVT transfer modes: e.g. de Sousa et al. (2011), Läubli et al. (2013) or Valor Miró et al. (2015) (for more information see the second article of this PhD, chapter 3). The most common studies on this subject matter are research works on the quality assessment of machine translated of post-edited subtitles (e.g. Armstrong et al., 2006; Volk, 2008; or Bywood et al., 2013), a topic that will be discussed in the following subsection and on the last two articles of this PhD.

1.2.3. *Quality and Quality Assessment*

Quality has been a concern for Translation Studies researchers since the very beginning of the discipline back in 1960's, and many researchers have dealt with its assessment (e.g. Nida, 1964; Carroll, 1966; Koller, 1987; Toury, 1995; Gambier, 1998; House, 2006; Melby, 2006; Hansen, 2008; Hague et al., 2011; or Fields et al., 2014) and ISO has already published standards for translation services, being ISO17100:2015, which is currently under revision and further development (ISO17100:2015/CD Amd1), the last one. The ISO17100 presents the standards that translation service providers (TSP) should follow in order to deliver a quality translation service following their particular set of specifications. It hence deals with the core processes, the resources and any other aspects –e.g. industry codes, best-practice guides– and legislation, that are needed to achieve such quality translation service. However, this ISO norm does not include any quality standards for the use of raw output of MT engines or the use of PE MT.

According to House (2006), quality assessment (QA) is product-based, and the QA models vary depending on the translation theory that lies behind it, being the following the main approaches to QA:

(1) Psycho-social approaches:

1. **Anecdotal approaches:** The main characteristic of these approaches is that the authors, who are usually professional translators, seek to assess translations based on its faithfulness to the original text (e.g. Savoy, 1968). The judgments are based on impression, feeling, and how good or bad one finds a translation.

2. **Neo-hermeneutic approaches:** According to these approaches, the quality of a translation relies on how fully a translator identifies with the original text (e.g. Kupsch-Oseriet, 1994). They also base the quality of a translation on the understanding of it (e.g. Gadamer, 1960; Steiner, 1975). Neo-hermeneutic approaches link the quality to the translators and their interpretation of the original text because translators tend to deliver an "optimal

translation" based on their intuition, empathy, interpretative experience and knowledge.

(2) Response-based approaches:

1. **Response-oriented approaches:** These approaches are communicative oriented; the assessment is based on how intelligible and informative a translation is (e.g. Nida, 1964; Nida and Taber, 1969). A good translation is to be equivalent to the original text in terms of response manner (dynamic equivalence).
2. **Functionalism:** According to these approaches, the quality of a text depends on the purpose (*skopos*) and functionality of the translation (e.g. Reiss and Vermeer, 1984).

(3) Text and discourse-oriented approaches:

1. **Literary-based approaches:** These approaches consider that the quality of a translation depends on the form and function of a translation within the system of the target culture (e.g. Toury, 1995; Baker, 1992; or Krein-Kühle, 2014).
2. **Philosophical, socio-cultural and socio-political approaches:** According to these approaches the quality of a translation is directly linked to the level of manipulation because of power relations, injustices, etc. in the target text (e.g. Venuti, 1995; Robinson, 2004).
3. **Linguistic approaches:** These approaches base the quality of a translation on the relationship between text or textual features, and how they are perceived by authors, translators, and readers (e.g. Nida, 1964; Catford, 1965; or Hatim and Mason, 1997).

4. **Pragmatic linguistic approaches:** According to these approaches, the quality of a translation depends on the relative match of the linguistic-situational particularities of the source and target texts (e.g. House, 2006).

The different approaches propose various methods, such as quantitative and qualitative testing by competent judges, comparison of translations against reference models, sentence-rating according to pre-established scales of intelligibility and informativeness, or gathering of respondents' opinions.

Quality and quality assessment have also been addressed in the field of MT and PE (e.g. Hutchins and Somers, 1992; Krings, 2001; King et al., 2003; Fiederer et al., 2009; Armstrong et al., 2006; or Melby, 2014). In this field of Translation Studies, QA can be done by human judges and/or automatic measures. The approaches presented above have only been performed by human judges so far. Automatic measures compare and correlate translations or post-edited MT output with the translations produced by MT engines in order to set the quality of the machine translations. There are different types of automatic measures: BLEU (Papineni et al., 2011), H-BLEU (Snover et al. 2006), NIST (Snover et al., 2008), and METEOR (Lavie et al., 2009) are measures of precision and compare the MT output to reference translations or post-edited documents. TER (Snover et al., 2006), H-TER (Snover et al., 2006), WER, and PER (Tillmann et al., 1997) are editing-distance measures, which calculate the number of modifications the MT output needs in order to resemble a reference translation or post-editing.

Human-based evaluations have been also applied to raw MT output and PE (e.g. O'Brien, 2005; Aziz et al., 2012, Fiederer et al., 2009; Avramidis and Popovic, 2013; or Lommel et al., 2014). Although most of these human-based evaluations focus on the fidelity or accuracy of the translation, its intelligibility or clarity, and its style, some use PE as a measure of assessment or classify the errors produced by the MT engines.

Although research on QA of post-edited text has increased, it is still rather limited. Fiederer and O'Brien (2009), Plitt and Masselot (2010), Carl et al. (2011), García (2011), Guerberof Arenas (2009, 2012), Melby et al. (2014) and Mariana (2014) have dealt with quality in post-editing, to a greater or lesser extent. Up until now, QA has been based mostly on what has been has termed in the QTLauchPad project (Lommel et al., 2014) as:

(1) **Holistic approaches:** These approaches assess the quality taking into account the whole text. Examples of it are Plitt and Masselot (2010), Carl et al. (2011) or Fiederer and O'Brien (2009). More information on this topic can be found in the third and fourth articles of this PhD (see chapters 4 and 5).

(2) **Analytic approaches:** These approaches assess the quality of the text by analyzing the text in detail according to a specific set of specifications. These approaches were adopted by García (2011) or Guerberof Arenas (2009, 2012). More information on this topic can be found in the third and fourth articles of this PhD (see chapters 4 and 5).

(3) **Mixed approaches:** These approaches assess the quality using both, an assessment of a translation as a whole, and an assessment of the text in detail using a set of specifications. The approach used in this PhD is based on Melby et al.'s (2014), Mariana's (2014) and Lommel et al.'s (2014) Multidimensional Quality Metrics (MQM), an approach that provides a framework to define metrics and scores used to assess the quality of human translated, post-edited, or machine translated texts. It also presents error categories, otherwise called issue types, which are used to assess different aspects of the quality and identify problems. This dissertation contributes in this field by including specifications needed to assess AVT oral modalities. More information

on this topic can be found in the third and fourth articles of this PhD (see chapters 4 and 5).

In the specific field of audiovisual translation, research on post-editing quality assessment is quite limited, as it has been researched mainly in EU-financed project such as SUMAT (Etchegoyhen et al., 2014), and e.g. Armstrong et al. (2006), or Aziz et al. (2012). In the field of AVT, some other studies on quality have also been carried out (e.g. Gambier, 2008; Chiaro, 2008).

Ultimately, this PhD finds its basis, on the one hand, on Franco et al. (2010) portrait of VO and OD and Matamala (2009b) description of the translation of wildlife documentaries. On the other hand, it is based on Kring's (2001) division of PE effort, O'Brien's (2006) study on how to measure PE temporal and technical efforts and, Lacruz et al.'s (2014) measure for PE cognitive effort. Finally, this PhD finds its grounds on MQM and Melby et al. (2014) but also on the holistic approaches to QA by Plitt and Masselot (2010), Carl et al. (2011), Fiederer and O'Brien (2009), as well as analytic approaches by García (2011) and Guerberof Arenas (2009, 2012).

1.3. Methodology

The multidisciplinary nature of this research work also affects its methodology, as it varies depending on the objectives to be fulfilled and the hypotheses to be validated in each of the articles. In this section, a general description of the methodology used is presented chronologically in order to clarify the process that has been followed throughout the PhD, and an indication of the article or articles where these methodological tools have been used is added. Two main approaches have been taken: on the one hand, a descriptive-analytical approach on the first stages of the thesis, based on a corpus created *ad hoc*; on the other, an experimental approach including two research studies –one on PE effort and another on QA– and a reception study.

(1) *Bibliographical review*

A bibliographical review on the characteristics of VO and the translation of documentaries through VO and OD was carried out in order to find possible challenges to appear when applying MT and PE MT into the current process of translation of documentaries. Furthermore, another bibliographical review on experiments studying post-editing effort and quality assessment was fulfilled to set the basis of the experimental studies that are part of this PhD.

(2) *Corpora creation*

Thanks to the collaboration of professional translators, documentary scripts in English (108, originals) and Spanish (92, translations) were collected in order to determine the characteristics of original scripts in English and translated scripts in Spanish. They were processed and analysed during a research stay in the translation company Pangeanic (Valencia), which was extremely valuable, as it helped not only with the analysis and process of the scripts but also with a better and more professional understanding of MT engines. The scripts, which contained either only a narrator –to be off-screen dubbed– or a narrator with interviewers, interviewees and/or spontaneous speech –to be voiced-over–, were divided in three different subcorpora:

- corpus of documentary scripts in English (En-DOC corpus);
- corpus of documentary scripts in Spanish (Spa-DOC corpus);
- corpus of randomly selected segments from the first corpora along with their human translation in Spanish and the translations produced by 8 MT engines –Apertium, Bing, Google Translate, Lucy MT, Promt, Reverso, Systran, Yandex– (Bil-DOC corpus).

More information on these corpora can be found in the first article of this PhD (see chapter 2).

(3) *Corpora Analysis*

The corpora were analysed differently depending on their characteristics:

1. Both En-DOC and Spa-DOC corpora were firstly analysed according to their micro- and macro-structures independently (Van Dijk, 1973). A bottom-up approach was used for the macrostructure analysis, which is understood in this study as the layout or overall structures within a documentary. For the analysis of the microstructure, which is known to be in this research as the connections between words and sentences of a documentary that set the basis of its general meaning, a bottom down approach was chosen. Afterwards, the results of En-DOC and Spa-DOC corpora analyses were compared.

2. The Bil-DOC corpus was analysed in three different ways: Firstly, an automatic evaluation of the results using BLEU and TER automatic measures was carried out. Secondly, the output of the MT engines was subjectively assessed by the researcher, who marked all the errors and classified them according to an Error Typology based on Uszkoreit et al's (2013) MQM. Finally, the found errors were analysed and compared.

The previous three steps correspond to the methodology used in the first article presented in the PhD (see chapter 2). After this descriptive stage, an experimental approach with the following steps was taken:

(4) *Stimuli creation for the post-editing experiments*

The selection of the documentary film excerpts to be used for all the experiments that are part of this PhD was based on four requirements:

- ♦ They had to be a short documentary or part of a long documentary in English; as the longer the documentary, the longer the experiment would be and, time wise, a short experiment was required, as the participants were volunteers.

- They had to have both narrator and experts talking in order to represent, as acurately as possible and only in one excerpt, as many different types of speakers that can be found in documentary films.

- They had to be as similar as possible to be comparable. Hence, they had to have the same amount of lines of dialogue for the narrator and the experts. They also had to be about the same length both in terms of number of words and minutes.

- They had to be able to be understood on their own. Therefore, they were edited to tell self-contained stories.

Finally, a part of a documentary film that fulfilled the requirements was found: *Must Watch: a Lioness Adopts a Baby Antelope* <https://www.youtube.com/watch?v=m2w-1BfHFKM>, a 7-minute excerpt of the episode *Odd Couples* from the series *Unlikely Animal Friends* broadcasted in 2009 by National Geographic. From that excerpt, two smaller excerpts with the following characteristics were selected:

		1st Excerpt	**2nd Excerpt**
Duration		101 seconds (1:41 minutes)	112 seconds (1:52 minutes)
Number of words		283	287
Lines of Dialogue	***Narrator***	5	4
	Female expert	3	3
	Male expert	1	1
	Inserts	0	1

Table 2. *Excerpts' characteristics.*

Both excerpts were machine translated from English into Spanish using Google Translate MT Engine, as it was the best MT engine according to the analysis of the Bil-DOC corpus analysis.

(5) *Experiment on PE effort*

This experiment intended to compare the effort of post-editing with the effort of translating in order to determine whether the PE effort is lower than the translation effort. The methodology followed in this experiment is described in the second article contained in this PhD (see chapter 3).

(6) *Experiment on PE quality*

This experiment intended to compare the quality of post-edited MT output with the quality of translations in order to prove whether the quality of the post-edited texts is equal to the quality of the translations according to evaluations made by experts of the field. The methodology used in this experiment is described in the third article of this PhD (see chapter 4).

(7) *Stimuli creation for the reception study*

The best translations and PEs, according to the QA carried out in the previous experiment, were recorded by voice talents at the Escola Catalana de Doblatge in order to be used for the reception study. Furthermore, observational notes were taken by the researcher and analysed afterwards as another way of evaluating the excerpts. The methodological approach is described in detail in the fourth article contained in this PhD (see chapter 5).

(8) *Experiment on user reception*

This experiment also aimed to compare the quality of post-edited MT output with the quality of translations. However, the evaluation was not only made by experts of the field, but also by the dubbing studio where the excerpts were produced, and the end-users. The methodology used in this experiment is described in the fourth article of this PhD (see chapter 5).

In the following table, the correlation between the steps followed for the methodology and the article for which they were used are summarized:

Art. 1	Art. 2	Art. 3	Art. 4	Methodology: steps
√				Bibliographical review
√				Corpora creation
√				Corpora analysis
	√	√	√	Stimuli creation
	√			Experiment on PE effort
		√	√	Experiment on PE quality
			√	Stimuli creation for the reception study
			√	Experiment on user reception

Table 3. *Methodology: steps*

1.4. PhD Structure

This PhD is presented as a compendium of publications and contains all the elements the regulations require for a PhD by compendium of publications: an introduction, the articles, a summary of the articles both in English and Spanish, a discussion of the obtained results and the conclusions. Furthermore, a bibliography and four annexes are included. The articles are presented according to a linear and chronological structure.

The first part of the PhD contains a descriptive chapter:

In **Chapter I**, the structure of the PhD and its objectives and hypotheses are presented. It also offers a brief state of the art for each of the subjects covered in this PhD, focusing on the works that set the theoretical framework of this research work. This section has been added in order to unify the theory that builds the frame of this research and set it within the current outlook. More specifically, it deals with previous work on VO and OD and on the translation of documentaries. Secondly, it summarizes the main research done in the field of

MT focusing on AVT and on PE effort. Finally, it provides an overview of academic works on QA. This summary is brief in nature and its aim is to contextualize our research in a wider framework.

The second part is divided in four chapters, each of them containing one of the four articles that form this PhD:

- **Chapter II** contains the first article of the PhD.

Ortiz-Boix, C. (Forthcoming). Post-Editing Wildlife Documentaries: Challenges and Possible Solutions. *Hermeneus*. 18.

In this article the challenges of introducing MT into the process of translation of documentaries for VO and OD are explored. Furthermore, possible solutions to the presented challenges are discussed.

- In **Chapter III,** the second article of the PhD is presented.

Ortiz-Boix, C. and Matamala, A. (2016). Post-Editing Wildlife Documentary Films: a New Possible Scenario*? Journal of Specialized Translation (JosTrans).* 26, 187-210.

In this article the possibility of including MT and PE into the process of translation of documentaries for VO and OD in terms of effort is explored.

- **Chapter IV** contains the third article of the PhD.

Ortiz-Boix, C. and Matamala, A. (Forthcoming). Assessing the Quality of Post-Edited Wildlife Documentaries. *Perspectives. Studies in Translatology*.

This article studies the quality of post-edited in comparison to the quality of translated documentaries by analyzing the assessment carried out by professionals on the field and VO lecturers in MA studies.

- In **Chapter V,** the fourth and last article of the PhD is presented.

Ortiz-Boix, C. and Matamala, A. (2015). Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: a Three-Level Approach. In: O'Brien, S. and Simard, M. (Eds). *Proceedings of the 4th Workshop on Post-Editing*

Technology and Practice (WPTP4). 16-30. Available at: <http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf>

The last article also explores the quality of post-edited documentaries. However, the QA presented here takes into consideration not only the assessment carried out by professionals on the field and VO lecturers in MA courses but also an assessment by a dubbing studio and end-users.

The third and last part of the PhD comprises two chapters:

- **Chapter VI** contains a summary of the PhD, as well as a summary of the results displayed throughout the PhD.

- In **Chapter VII,** the results are discussed and the conclusions of the PhD are presented. Future research venues are also put forward.

Afterwards, the updated bibliography, the filmography, and four annexes are presented. An updated bibliography is added as the articles are included in their original form and some forthcoming references may already be published. The annexes are presented as follows.

(1) Annex A (in paper) contains the edited versions of the articles, in their format at the moment of presentation of this dissertation, according to the journals constrains.

(2) Annex B (electronic) includes the documents used during the experiment on PE effort presented in the second article of this PhD:

- Documentary *Must Watch: a Lioness Adopts a Baby Antelope* (whole video).

- Excerpt 1 of the documentary *Must Watch: a Lioness Adopts a Baby Antelope* (video).

- Excerpt 2 of the documentary *Must Watch: a Lioness Adopts a Baby Antelope* (video).

- ♦ Transcripts of excerpt 1 and excerpt 2.
- ♦ Machine Translations of excerpt 1 and excerpt 2.
- ♦ Questionnaire pre-task both in their original language and in English.
- ♦ Questionnaire post-task both in their original language and in English.
- ♦ Instructions both in their original language and in English.
- ♦ Informed consent form and Informative document for the experiment on PE effort, both in their original language and in English.

(3) Annex C (electronic) consists of the documents used during the experiment on QA by experts presented in the third article of this PhD:

- ♦ Documentary *Must Watch: a Lioness Adopts a Baby Antelope* (whole video).
- ♦ Excerpt 1 of the documentary *Must Watch: a Lioness Adopts a Baby Antelope* (video).
- ♦ Excerpt 2 of the documentary *Must Watch: a Lioness Adopts a Baby Antelope* (video).
- ♦ 12 documents with the transcript of excerpt 1 + its translation.
- ♦ 12 documents with the transcript of excerpt 1 + its post-editing.
- ♦ 12 documents with the transcript of excerpt 2 + its translation.
- ♦ 12 documents with the transcript of excerpt 2 + its post-editing.
- ♦ Questionnaire pre-task both in their original language and in English.
- ♦ 2 questionnaires post-task both in their original language and in English.

- Instructions both in their original language and in English.

- Informed consent form and Informative document for the experiment on QA by expert and the user reception study, both in their original language and in English, both in their original language and in English.

(4) Annex D (electronic) contains the documents used during the user reception study presented in the fourth article of this PhD:

- Documentary *Must Watch: a Lioness Adopts a Baby Antelope* (whole video).

- Excerpt 1 of the documentary *Must Watch: a Lioness Adopts a Baby Antelope* (video).

- Excerpt 2 of the documentary *Must Watch: a Lioness Adopts a Baby Antelope* (video).

- 12 documents with the transcript of excerpt 1 + its translation.

- 12 documents with the transcript of excerpt 1 + its post-editing.

- 12 documents with the transcript of excerpt 2 + its translation.

- 12 documents with the transcript of excerpt 2 + its post-editing.

- Questionnaire pre-task both in their original language and in English.

- 2 questionnaires post-task both in their original language and in English.

- Instructions both in their original language and in English.

- Informed consent form and Informative document for the user reception study, both in their original language and in English.

- Edited version of the best translation of excerpt 1.
- Edited version of the best translation of excerpt 2.
- Edited version of the best post-editing of excerpt 1.
- Edited version of the best post-editing of excerpt 2.
- Translated VO version of excerpt 1 (video).
- Post-edited VO version of excerpt 1 (video).
- Translated VO version of excerpt 2 (video).
- Post-edited VO version of excerpt 2 (video).

Chapter 2. First Article

2. First Article

Ortiz-Boix, C. (Forthcoming). Machine Translation and Post-Editing in Wildlife Documentaries: Challenges and Possible Solutions. *Hermeneus*. 18.

ABSTRACT: This article presents some of the challenges that may have to be overcome in order to introduce machine translation (MT) into the process of translating wildlife documentary films. Until now, MT has mainly been applied to written general and specialized texts. However, in the past few years, EU-financed projects have started to work in the field of audiovisual translation with the aim to introduce MT into subtitling. It has already been proven that post-edited machine translated subtitles can reach the appropriate quality levels. Nevertheless, in the case of documentaries, not only subtitling but also voice-over and off-screen dubbing can be found in countries where subtitling is not the main audiovisual transfer mode. Therefore, similar research in voice-over and off-screen dubbing is believed to be worthy. This article aims to describe the challenges of machine translating documentary scripts by presenting a preliminary analysis on the translations produced by MT engines. Firstly, an overview of the characteristics of voice-over and off-screen dubbing is provided, as well as a brief review dealing with MT and post-editing in audiovisual translation. Next, the methodology used to carry out the analysis of both a corpus of documentary scripts and a corpus of machine translations of documentary scripts is explained. Finally, before summarizing new potential avenues of research, the challenges that may have to be faced in order to achieve high quality translations of documentary scripts using MT are pointed out, the results of the analysis are presented, and some possible solutions are suggested.

KEY WORDS: post-editing, audiovisual translation, voice-over, off-screen dubbing, documentaries, machine translation, pre-editing.

1. Introduction

Research on machine translation (MT) and post-editing (PE) has attracted great interest over the last decade, not only among Translation Studies scholars, but also among translation industry stakeholders. TAUS (Joscelyne, 2009) market study indicates that 92.23% of the language server providers included in its study already use or intend to use MT and PE as part of their translation process. However, in the audiovisual translation (AVT) market, professional experiences in MT and PE are limited (Volk et al., 2010) and industry voices in favour of MT are just beginning to be heard (Georgakopoulou, 2010). Interest in academia has increased in recent years, focussing on the implementation of MT and PE in subtitling, in part due to EU-financed projects such as eTITLE (Melero et al., 2006), EU-Bridge (Waibel, 2012), or SUMAT (Del Pozo et al., 2012). The promising results of these studies (Fishel 2012; Bywood et al., 2013; Freitag et al., 2013) have encouraged other researchers to study the inclusion of MT in other AVT modes such as audio description (Ortiz-Boix, 2012; Fernández et al., 2013).

Inspired by existing research, I have started an investigation based on the hypothesis that MT can be successfully implemented when translating wildlife documentaries for oral transfer modes such as voice-over (VO) and off-screen dubbing. This research will assess the quality of MT output, and most importantly, the PE effort as compared to a standard human translation. However, before carrying out this experimental part of the research, I have considered it relevant to do a bibliographical survey and carry out a qualitative analysis on a corpus of documentaries, in order to point out the specific problems that will probably have to be addressed. As documentary films can deal with a wide variety of subjects, such as arts, health, history, music or wildlife, to mention but a few, and each topic has its own terminological specificities, a specific domain has been selected to narrow down the analysis: wildlife. This is due to the fact that there is a wide variety of wildlife documentary films; while some present species or ecosystems through beautiful images, the voice of a narrator, and sometimes, of experts (*Planet Earth*, 2006), others are almost reality programs (*The Crocodile Hunter*, 1997-2004). Furthermore, wildlife documentaries are frequent in TV's daily schedule – both in Spanish and English-speaking countries, illustratively on channels such as Animal Planet, BBC One, BBC Four, National Geographic Wild, La 2, and Canal Plus. The article aims to present the potential challenges arising from the use of MT engines and PE software when translating for VO and off-screen dubbing, two audiovisual transfer modes which can be often found in wildlife documentary films.

The article focuses on eight challenges and their possible solutions: (1) spotting, (2) synchronization, (3) access to audiovisual content, (4) variety on the script format, (5) register variety within a same script, (6) terminology, (7) errors and inaccuracies in the original script, and (8) linguistic inconsistencies in the original script. In order to identify the challenges, two approaches have been taken: on the one hand a bibliographical survey of existing literature on VO and off-screen dubbing has been conducted, and on the other, an analysis of two corpora, namely a corpus of wildlife documentary scripts in English and a corpus in Spanish. An error analysis of a corpus of 50 sentences machine translated using eight free online engines provides additional insight into the most common errors produced by MT engines.

The article is divided as follows: a short overview on the two transfer modes under analysis (VO and off-screen dubbing), as applied to the translation of documentaries, as well as a short review of previous MT and PE research within AVT are presented in sections 2 and 3 respectively. In section 4, the methodology used to identify the challenges is explained. Sections 5, 6 and 7 present the challenges: section 5 focuses on the challenges found in previous academic works, section 6 describes those derived from the analysis of corpora 1 and 2, and section 7 lists the challenges found through both the automatic and human evaluations of the corpus of 50 sentences. In section 8, possible solutions are proposed, and in the last section, conclusions and further research are presented.

2. Voice-over and Off-screen Dubbing

The branch of Translation Studies that deals with documentary films is AVT, which can be described as the field of Translation Studies concerned with the transfer of multimodal and multimedia texts into another language and/or culture (Baldry and Thibault, 2006). Although there are many AVT transfer modes (subtitling, dubbing, audio description, surtitling, voice-over, subtitling for the deaf and hard of hearing, live subtitling, video-game localization, etc. [Remael, 2010]) and almost all of them could be used in a documentary film, this article focuses only on off-screen dubbing and VO of wildlife documentary films, from English into Spanish. These two modes have been selected as they are the most used in open and closed TV channels in Spain, for instance, where it is common to find documentaries in which the narrator is re-voiced by using off-screen dubbing, whilst interviewees are rendered via VO. Although research in these transfer modes and genres initially received little attention, the trend has changed in recent years with some more works being published: Espasa (2004), Franco (2000, 2001a, 2001b), García Luque (2011), Matamala (2002, 2004, 2008, 2009a, 2009b), and Orero (2004, 2007).

Díaz Cintas and Orero (2006: 473) define voice-over as follows:

> Technique in which a voice offering a translation in a given target language is heard simultaneously on top of the SL voice. As far as the soundtrack of the original program is concerned, the volume is reduced to a low level that can still be heard in the background when the translation is being read. It is common practice to allow the viewer to hear the original speech in the foreign language at the onset of the speech and to reduce subsequently the volume of the original so that the translated speech can be inserted. The translation usually finishes several seconds before the foreign language speech does, the sound of the original is raised again to a normal volume and the viewer can hear once more the original speech.

According to Franco et al. (2010: 25), voice-over translation in factual programmes is said to help reproduce the feeling of reality, truth and authenticity that the original audiovisual product gives, which is supported both by visual evidence (images of events, people, documents and archival footage) and by verbal evidence (interviews with experts and witnesses). The delivery of VO does not usually show regional accents in the target text and does not generally reproduce specific oral features such as fluffs, hesitations or grammatical mistakes. Orero (2006) highlights the importance of three types of synchrony in VO: kinetic synchrony –the voice delivering the translation matches the body movements which can be seen on screen–, action synchrony –the voice delivering the translation matches the actions taking place on screen–, and voice-over isochrony –the translated message fits between the beginning and the end of the original speech, leaving some time before it starts and after it ends during which the original soundtrack is heard.

Off-screen dubbing, also termed commentary and narration by authors like Pönniö (1995), shares kinetic and action synchrony with voice-over but not voice-over isochrony. This is because the original voice is not heard but instead substituted by the target language. Additionally, VO is generally used for semi-spontaneous or spontaneous interviewees, whilst off-screen dubbing is usually applied to narrators with a planned discourse, and this also has implications in the language register.

Other shared features pointed out in the literature (Franco et al., 2010) are the lack of postproduction scripts or, if available, the poor quality of the transcriptions provided to the translators, which may contain linguistic errors and inaccuracies, etc. (see sections 6.2.4 and 6.2.5). Furthermore, as Matamala (2010) states, wildlife and scientific documentaries –the specific focus of this research– make use of a vast array of terminology, which might be a challenge for their translation (see section 6.2.3).

3. Machine Translation in Audiovisual Translation

So far, implementing MT into the translation process has proven successful in limited domains, such as meteorology or finances, and when working with general texts, in which case MT is used for gisting purposes and for interpersonal communication (Ray, 2004: 8-9). MT engines are becoming more and more domain-specific, which guarantees a better quality translation for the post-editors to work with (Läubli et al., 2013: 2). In the case of AVT, the implementation of MT is falling behind, as it has only been researched in subtitling (Melero, 2006; Armstrong et al., 2006; Volk, 2008; Bywood, 2013), and to a far lesser extent, audio description (Ortiz-Boix, 2012; Fernández et al., 2013).

Different approaches have been adopted to implement MT in the field of subtitling. Armstrong et al. (2006) have researched quality improvement when translating subtitles in the language pair English <> German with an EBMT engine with homogeneous data in comparison with an EBMT with heterogeneous data. The completed eTITLE Project (Melero, 2006) intended to increase the efficiency of subtitling by automating various processes within its workflow, achieving a good BLEU score (36.9) in the English-Spanish combination.

Research in this field has also been carried out by Volk (2008), who has investigated whether it is feasible to use MT in subtitling by focusing on the language combination Danish <> English and checking three criteria: number of users, customer satisfaction, and long-term usage of the MT system. He concludes that it is feasible as the statistical MT based system reached high BLEU scores (average 57.3) and saved time in the translation process. Furthermore, he points out the possibility of adding pre-editing to control the language of the source documents so that the MT system is more competitive.

In the case of audio description, Ortiz-Boix (2012) presents a preliminary study on the application of MT to audio description process in the Catalan <> Spanish language pair. Although it is a preliminary study within the context of an MA dissertation, the first results are reassuring as the lowest BLEU score was 67.00. MT is envisaged by this researcher as a tool to increase accessibility in multilingual environments by working with closely related languages (Matamala et al., forthcoming).

Finally, the most recent project on the topic, SUMAT (Online Service for Subtitling by MT, see http://www.sumat-project.eu/), works with 14 different language pairs and initial BLEU results of 25.5 are promising (Bywood, 2013). The project aims to provide not only automatic measures but also to test the human PE effort, an approach taken in general translation (De Almeida et al., 2010) but almost absent in AVT (Sousa et al., 2011).

To sum up, the existing results regarding the application of MT and PE to subtitling and audio description processes have compelled us to put forward the hypothesis that MT with PE could also be successfully implemented into the translation of documentary films. Before carrying out experimental research to prove this hypothesis, a qualitative analysis has been done to foresee possible challenges, as described in the next section.

4. Methodological Considerations

Two methodological approaches have been adopted: on the one hand, a bibliographical review, which has led us to identify three challenges (discussed in section 5), and on the other, an analysis of three corpora which has allowed to confirm some of the issues found in the bibliographical survey, and to add some new ones (see sections 6 and 7). The main features of the corpora and how they have been analysed are explained next.

4.1. Corpus Creation

In order to find the characteristics of documentary scripts that can impact MT and PE processes, 108 documentary scripts in English (original texts) and 92 in Spanish (translations) have been collected and analysed. Some of the documentaries (66) only contain a narrator to be revoiced using off-screen dubbing, whilst others (54) contain a narrator plus interviewees and spontaneous speech to be voiced-over. These scripts were divided into three corpora:

(1) *En-Doc:* 108 English documentary scripts in English, containing 504,368 words in 13,426 sentences (see table 1).

(2) *Spa-Doc:* 92 documentary scripts in Spanish containing 440,651 words in 7,053 sentences. 80 of them are human translations of the documentaries included in En-Doc, whilst the remaining 12 are also human translations whose original script is not included in the previous corpus (see table 1).

CORPUS	SCRIPTS	SEGMENTS	WORDS
En-Doc	108	13,426	504,368
Spa-Doc	92	7,053	440,651

Table 1. *En-Doc & Spa-Doc Corpora*

(3) *Bil-Doc:* constituted by a random selection of 50 original English segments (meaning group of words, i.e. whole sentences or syntagmas the MT engine is fed with) next to their human translation and eight MTs into Spanish. It contains 6,592 words (633 English and 5,959 Spanish words), as shown in table 2:

CORPUS	LANGUAGE	SEGMENTS	WORDS
Bil-Doc	English	50	633
	Spanish	450	5,959

Table 2. *En-Doc & Spa-Doc Corpora*

The 50 random segments in English and their translations in Spanish were extracted from the 80 documentary scripts the Spanish translation of which was already available. Only text that has to be voiced –and therefore needs to be translated– was considered and additional information on the visuals or music –generally omitted from the translation but sometimes included in the scripts– was disregarded in this selection. The segments were translated using the English into Spanish free online MT engines that were found on the web, (of which there are only eight) when the analysis took place (table 3):

MT Engine	Website
Apertium	www.apertium.org/#translation
Bing	www.bing.com/translator
Google Translate	http://translate.google.com/
Lucy MT	www.lucysoftware.com/english/machine-translation/lucy-lt-kwik-translator-/
Promt	www.online-translator.com/
Reverso	www.reverso.net/text_translation.aspx?lang=ES
Systran	www.reverso.net/text_translation.aspx?lang=ES
Yandex	https://translate.yandex.com/

Table 3. *MT Engines*

4.2. Corpus Analysis

En-Doc and Spa-Doc corpora helped determine some of the challenges regarding wildlife documentary scripts' features. Both micro- and macro-structures of documentaries in English and Spanish were analysed and compared. Macro-structures are "the overall structures of a text" (Van Dijk, 1973: 73), whilst micro-structures are understood as the connections between words and sentences within a text which become the basis for its general meaning (Van Dijk et al., 1983: 73).

(1) *Macro-structure analysis in both En-Doc and Spa-Doc corpora*: a manual analysis of the script layout was carried out and divergences were found in the formatting of time codes and the inclusion of additional contents (description of visual information, details about the music heard, etc.). The results of this corpus-based bottom-up analysis, which was not based on any previous categorisation, were compared with the script layouts found in Franco et al. (2010). This analysis, the results of which can be found in sections 5.1 and 6.1, was carried out for both En-Doc and Spa-Doc corpus independently, and the results were then compared.

(2) *Micro-structure analysis in both En-Doc and Spa-Doc corpora:* this analysis adopted a different approach, resting on a pre-established categorisation from previous literature. A list of categories (namely terminology, register, linguistic inconsistencies, inaccuracies and errors in the original script) was searched manually in the corpus in order to confirm or reject their presence, hence offering qualitative data through a top-down corpus-based analysis. This analysis, the results of which can be found in sections 5.2, 6.2, 6.3, 6.4, and 6.5, was carried out for both En-Doc and Spa-Doc corpora independently, and then results were compared.

(3) *Analysis of the Bil-Doc corpus:* this corpus was used to confirm some of the previously found challenges regarding micro-structure, as well as to run a preliminary test on the possible application of MT to wildlife documentary films, and to determine the most common errors when machine translating wildlife documentary films. Therefore, an automatic and a human subjective evaluation were made.

In order to analyse the Bil-Doc corpus and to evaluate the translations, several steps were followed:

(1) *An automatic evaluation,* the results of which can be found in section 7.1, was made using Asia Online software (www.asiaonline.net), providing BLEU and TER automatic measures of the eight MT engines' translations against the existing human translations.

(2) *A subjective assessment of the output from all eight MT engines* was made by one researcher (results can be found in Section 7.2). All errors were marked and classified according to a table based on the Multidimensional Quality Metrics Error Typology (MQM) proposed by Uszkoreit et al. (2013). Quality assessment of human translations

has been researched by many authors in translation studies (e.g. Hurtado Albir, 2001; Williams, 2001; Eckersley, 2002; Hurtado Albir, 2007; Nord, 2014), who have proposed different categorizations of errors. However, they do not take into account the specificities of MT. This is why a categorization of errors specifically for MT output was considered the most appropriate for the presented assessment, as the analysed output was machine translated. Among all error categorizations available that asses MT output (e.g. Font Llitjós et al., 2004a; Koponen, 2010), MQM was selected as a starting point because it is the most exhaustive and allows researchers to introduce domain-specific categories or erase unneeded categories. In any case, only categories regarding accuracy, issue, type and mechanical issues included in fluency were used for the purposes of this article as they are considered the most relevant (Uszkoreit et al., 2013). Table 4 lists all error categories used in this article:

ACCURACY	Terminology		A term is translated with a term other than the one expected for the domain or otherwise specified.
	Mistranslation		The target content does not accurately represent the source content.
		Overly Literal	The translation is overly literal
		False Friend	The translation has incorrectly used a word that is superficially similar to the source word.
		Sould not have been translated	Text was translated that should have been left untranslated.
		Date/time	Dates or times do not match between source and target.
		Unit conversion	The target text has not converted numeric values as needed to adjust for different units.
		Number	Numbers are inconsistent between source and target.
		Entity	Names, places or other "named entities" do not match.
	Omission		Content is missing from the translation that is present in the source.
	Addition		The target text includes text not present in the source.
	Untranslated		Content that should have been translated has been left untranslated.
FLU	Spelling		Issues related to spelling of words.
		Capitalization	Issues related to capitalization.
		Diacritics	Issues related to the use of diacritics
	Typography		Issues related to the mechanical presentation of text. The category should be used for any typographical errors other than spelling.
		Punctuation	Punctuation is used incorrectly for the locale or style.
		Unpaired quote marks or brackets	One of a pair of quotes or brackets is missing from the text.
	Grammar		Issues related to the grammar or syntax of the text, other than spelling and orthography.

E N C Y		Morphology	There is a problema in the internal construction of a word.
		Part of speech	A word is the wrong part of speech.
		Agreement	Two or more words do not agree with respect to case, number, person or other grammatial features.
		Word order	The word order is incorrect.
		Function words	A function word is used incorrectly.
	Unintelligible	The exact nature of the error cannot be determined. Indicates a major break down in fluency.	

Table 4. *Used Metrics for human evaluation based on MQM (Uszkoreit et al. 2013)*

After categorising the errors by marking and processing them with an Excel spreadsheet, the results of each MT engine were analysed and compared.

Before introducing the results of the analyses, namely the foreseen challenges if MT is included in the process of translating wildlife documentaries to be voiced-over and off-screen dubbed, a summary of the methodology –including in which Section the results can be found– is presented in table 5.

Approach	**Corpus**	**Type of Analysis**	**Results in…**
(a) Bibliographical review			Section 5
(b) Corpus analysis	EN-DOC corpus	Micro- and macro-structure analysis	Section 6
	SPA-DOC corpus	Micro- and macro-structure analysis	Section 6
	BIL-DOC corpus	Automatic analysis	Section 7
		Subjective assessment	Section 7
		Final comparison	Section 7

Table 5. *Review of the methodology*

5. Challenges Based on Bibliographical Review

The bibliographical review has allowed us to identify three fundamental challenges, which are dealt with in this section: spotting, synchronisation, and access to the audiovisual content.

Synchronisation is a key feature of both voice-over and off-screen dubbing. Synchronisation is reached thanks to the careful work of audiovisual translators, who rephrase, condense or adapt the text so as to match the images and the time slots available. Moreover, to facilitate the recording by the voice talent, time codes are also included in their script, a task called spotting. Should MT be implemented in the working flow, a specificity would be that translators (or post-editors) would not only correct possible MT errors, but also adapt the text so as to comply with the various types of synchronies (Orero, 2006). Ideally, this would require a PE software which displays the audiovisual content and not only the written text.

5.1. Spotting

Spotting, also called timing or cueing, is the process of defining in and sometimes out time codes of each voice-over or off-screen dubbing unit. As stated by Díaz-Cintas and Remael (2007: 94), time codes are an essential tool, not only for subtitling, but also for the rest of AVT modes such as dubbing and voice-over. Spotting can be done by an audiovisual translator or by another professional, as it is also the case in subtitling (Sánchez, 2004), either before or after the translation. Various scenarios can be found in the profession: (1) the translator is given an already created spotting list, which is the case of templates (Sánchez, 2004; Díaz Cintas et al., 2007; Kapsaskis, 2011; Artegiani et al., 2014); (2) the translator is required to do the spotting and decide the time codes; or (3) the translator produces a translation without time codes and another professional does the spotting afterwards. In the second and the third scenarios, the ones considered by Franco et al. (2010) in their seminal book on voice-over, it is often the case that translators are given a transcript which includes time codes which do not correspond to the timing of the actual audiovisual content they receive. In the En-Doc corpus, scripts with and without time codes can be found, as illustrated in tables 6 and 7.

25m up in the treetops, old king Zog keeps everything in order...
His kingdom of leaves and branches rises above the Pantanal, the largest wetland in the world, and when the rainy season returns and the floodplains are submerged, his tree becomes a kind of island.
This marsh is so large that the only ones who really know where its boundaries lie are the migrating birds, who leave when it once again becomes dry and yellow.

Table 6. *Spotting. En-Doc. No Time Codes*

02;15 Kala's father and mother spent the winter on Hudson Bay. Each on its own, they trailed polar bears on the pack ice, feeding on the remains of seals left behind by the bears.
02;28 Before the end of the season, they returned to the tundra, mated and after 52 days of gestation, the female gave birth to her young.
02;43 For the first two weeks of her pups' lives, she had to stay with them deep in the den without ever coming out. At birth, they were blind and weighed only 50 grams each.

Table 7. *Spotting. En-Doc. Time Codes*

However, all translated scripts in our corpus contain time codes (see table 8), which not always coincide with the time codes in the original script (compare, for instance, the Spanish spotting in table 8 which corresponds to the original in table 7). Thus, translators needed to either introduce the spotting when translating the script or check and rewrite the time codes because they were different.

02:15
El padre y la madre de Kala pasaron el invierno en la bahía de Hudson. Cada uno por su lado, siguieron el rastro de los osos polares en la banquisa, alimentándose de los restos de focas que los osos dejaban atrás.

02:30
Antes de que terminara la estación, regresaron a la tundra, se aparearon, y, tras cincuenta y dos días de gestación, la hembra dio a luz a sus crías.

02:41
Durante las dos primeras semanas de vida de las crías, debía quedarse con ellas en el fondo de la madriguera, sin salir nunca de ella. Al nacer, las crías eran ciegas y pesaban solo cincuenta gramos cada una.

Table 8. *Spotting. Spa-Doc. Time Codes*

A specificity of voice-over and off-screen dubbing in the corpora and confirmed by the examples in Franco *et al.* (2010) is that, generally, only time codes in (and not out) are included.

An additional difference related to time codes is that in the English original scripts they appear in various formats whilst in the Spanish scripts –for voice-over and off-screen dubbing– the formatting is limited to two. This comes to show that, even in the uncommon scenario in which the time codes in the original script coincide with the target language time codes, adapting their format would be an additional requirement. As summarized in table 9, time codes within En-Doc corpus may indicate minutes and seconds (from type 1 to type 6); hours, minutes and seconds (types 7 to 10); hours, minutes, seconds and frames (from type 11 to 13) or feet (type 14). However, type 6 is the most commonly found among them. In the corpus Spa-Doc only two different time code formats are found: 00:01 (type 6) and 00.01 (type 5), the former being the most common one.

Type	**Time code**		**Type**	**Time Code**
1	(00.02)		8	01:00:10
2	01 08		9	10 04.06
3	0304		10	10.00.03
4	00;04		11	01:00:22:27
5	00.06		12	10 00 07 00
6	00:19		13	(01:08:18:00)
7	00.00.08		14	6.5

Table 9. *Types of Time Codes Spotting*

All in all, spotting is a must before a documentary is recorded. If MT with post-editing is implemented, dealing with the spotting might be a challenge, be it because time codes will have to be modified (if available) or included (if they do not appear in the original script). Therefore, introducing or correcting the time codes in the script, which will be fed into the MT engine, might be an adequate task to increase PE productivity.

5.2. Synchronization

The spotting or assignation of times codes can facilitate the synchronization of text and the audiovisual content according to the three types of synchronies to be reached when translating documentaries (Orero, 2006): kinetic synchrony, action synchrony, and isochrony. These synchronisations can only be achieved by confronting the actual translation to the audiovisual content, and in a scenario in which MT is implemented in the working flow, they may have to be carried out during the post-editing phase. However, some automatic strategies to reduce this load may be considered such as limiting the minimum and maximum number of characters per sentence, as already done, for example, by PET (Post-Editing Tool, see http://www.clg.wlv.ac.uk/projects/PET/), a post-editing research tool designed to help users post-edit and assess both MT output and human translations.

5.3. Access to Audiovisual Content

As Franco et al. (2010) state, the source text in AVT is the audiovisual product, which is made of images and audio. Scripts or transcripts, *i.e.* written texts, are sometimes provided to help the translator but it is not always the case. When machine translating, however, a written original text is needed, be it in the form of a pre-existing script, transcript, or automatic transcription of the audio. As visuals and audio are not considered in the automatic process, it is of the essence that the MT output is revised during the post-editing phase, not only in terms of language adequacy and fluency, but also in terms of written text-audiovisual content synchronisation. In order to do so, access to the visuals is needed, which, to the best of my knowledge, can only be achieved nowadays by using post-editing software plus video player. Available post-editing software, be it commercial CAT tools or applications for research purposes, do not allow rendering of audiovisual content in their interface. This is the case of PET (Aziz et al., 2012), CASMACAT (Ortiz-Martínez et al., 2012) or TCTool (Font Llitjós 2004b). Although SUMAT looks into the possible integration of MT with AVT, its platform and infrastructure does not integrate neither image nor audio (Del Pozo et al., 2013), which means that when carrying out SUMAT tests, participants had to work with standard subtitling software.

6. Challenges based on En-Doc and Spa-Doc Corpora Analyses

This analysis is based on the observation of En-DOC and Spa-DOC corpus and takes a closer look at some of the linguistic issues which affect either scripts' macro- or micro-structures, or both: variety on script format, register variety within the same script, terminology, errors and inaccuracies in the original script, and lexical problems in the original script.

6.1. Variety on the Script Format

As Franco et al. (2010) explain, original scripts formats provided to audiovisual translators differ substantially. After analysing the macro-structure (information contained within the scripts and how it is presented) of all the compiled scripts in the En-Doc corpus, several types of script layouts have been found. The obvious characteristic shared by all scripts is the transcription of narrations plus other speeches, from experts to spontaneous participants. However, it has been observed throughout the corpus that the transcription can be either included in a table which contains additional information or in a plain text document with nothing else but the time-codes.

When the script layout is presented in a table, narrations, also called commentaries, tend to be included under the heading *commentary*, or *comm*, whilst words from experts or spontaneous participants generally follow the term *sync*. It must be stressed that some scripts contain no differentiation between these two types of speakers, and when they appear together, they usually appear under the heading *audio*, *description*, *sync/comm* or *script*. Another feature of the table-based scripts comprised in the corpus is that time codes are always included, under the heading *time codes*, *time code, timecode* or *TC*. Many of these scripts also contain additional information, referring to elements such as images, music or even the mood of each character when talking, with varying degrees of detail. Two examples can be found in tables 10 and 11. Whilst the former indicates that the visuals correspond to boats on a river with no further details ("River-boats"), the latter describes more precisely what is seen ("Local people dancing &

playing instruments. Cuts to landscapes") and gives details as to the music that can be heard ("Siddhi Drumming").

TIME CODE	VISUALS	DIALOGUE/NARRATION
10 00 25	River – boats	In 1998, I left Italy and set off for the heart of Africa, to the Congo basin. The focus of my quest… lowland gorillas.

Table 10. *Variety of Scripts - En-Doc 1*

Timecode	In-Vision	Music	Sync	Narration
10.00.39	Local people dancing & playing instruments. Cuts to landscapes	10.00.44 Siddhi Drumming OUT		African features and rhythms, low thorny forests and the king of the beasts – all establish where we are – or does it?

Table 11. *Variety of Scripts - En-Doc 2*

On the other hand, and when the script layout is not presented in a table but in a basic text document, it only contains the transcription of the words with speech turns separated into paragraphs and with time codes at the beginning, if available (see table 12).

01 08 Butterflies are particularly well-known for their beautiful shapes and the splendid colours of their wings…

01 17 Their beauty has made them familiar to humans.

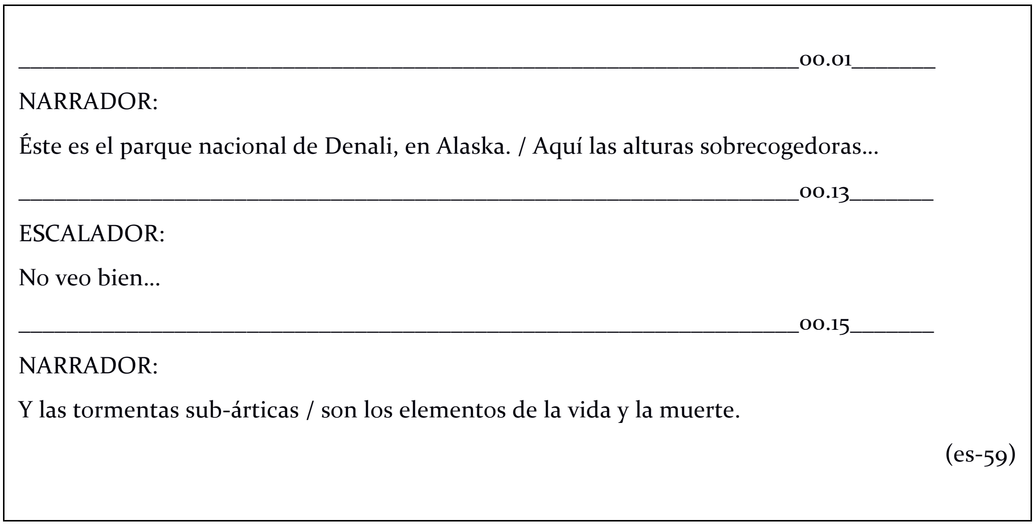
01 29 But butterflies are only part of a large family that we are not well acquainted with, the insects, the largest and most successful family of animals on planet Earth.

Table 12. *Variety of Scripts - En-Doc 3*

Despite the original English scripts can be presented in many different formats, the variety of script layouts in the case of their Spanish counterparts is not as large. Similarly to the original scripts, the translation of the scripts can be either presented in a table (see table 13) or in a plain text document (see table 14), which is the most common option. The latter option sometimes contains indications of the voice talents concerning the pauses to be made (see the slashes in table 14).

Chyros – TC'S	DECLARACIONES	NARRADOR
00.02.14		Antiguas leyendas de marineros hablan de islas misteriosas que se mueven empujadas por la corriente en un mar de tiempo.
02.26		Pueden aparecer y desaparecer de Nuevo en cualquier punto de la enorme extensión del océano. Y llevan el desastre a cualquiera que se acerque demasiado.

Table 13. *Variety of Scripts - Spa-Doc 1*



____00.01_______

NARRADOR:

Éste es el parque nacional de Denali, en Alaska. / Aquí las alturas sobrecogedoras…

____00.13_______

ESCALADOR:

No veo bien…

____00.15_______

NARRADOR:

Y las tormentas sub-árticas / son los elementos de la vida y la muerte.

(es-59)

Table 14. *Variety of scripts - Spa-Doc 2*

A correlation between the original script layout and the audiovisual transfer mode used in the translation can be found. The speeches, which are normally introduced by the word *narrator* or by no specific heading in the original script, correspond to a disembodied voice that is usually off-screen dubbed. They are generally transferred onto the translated script by indicating *narrador* (narrator) or nothing. The ones that are introduced by a specific proper name in the original script correspond to people talking on screen and are usually voiced-over. This is transferred onto the translated scripts by including the name of the on-screen speaker, a nick-name to identify the person, the symbol *VO* or the heading *declaraciones*. On occasions, a narrator or talking head may speak both on- and off-screen, in which cases, the symbols *sync* or *comm* are generally added to indicate whether they appear on- or off-screen in the original version.

All in all, two obvious but relevant conclusions for the use of MT should be highlighted: on the one hand, not all information contained in the original script is to be included in the translated version, and, on the other, translated script layouts are different from the original ones. This means that, most probably, an adapted translation script or template without all the extra information should be created before feeding the MT engine with it. Additional research is needed on how this additional task would impact the productivity and in which scenarios it would be worth it.

6.2. Variety of Registers within the same Script

While VO is used to translate the words of interviewed experts and spontaneous dialogue, generally on camera, off-screen dubbing is mostly used for narrators off-camera. Different speakers can coexist within a same wildlife documentary film, and depending on who is talking and the communicative situation, the register may vary:

(1) *Third person narrator*: as stated by León (1998: 18), "(t)he narrator-presenter plays a very important role in television documentary since his voice and statements to camera are the backbone in the structure of the programme." Narrators present and explain facts with the help of images, and sometimes, the presence of experts in the documentary. Their discourse is usually planned, based on a previously written script. In the corpus, their language is generally formal, although more colloquial or non-standard forms may appear occasionally, so as to engage the audience. See for instance, the rhetorical questions used to address the audience in table 15.

00:05
5 extraordinary stories from the wild.
00:08
But watch out because *there's* a twist. One of them is a *fake* created just to test you. Can you tell fact from fiction? Or will you be Fooled By Nature?
00:23
Nature's fantastic feeders.

Table 15. *Variety of registers - En-Doc 1*

(2) *First person narrator*: narrators may change from a third person commentary to a first-person in order to interact with other participants or to adopt a more subjective approach, as can be seen in table 16.

00:03 COMM Stephen Fry
Twenty years ago my good friend Douglas Adams spent a year tracking down endangered animals together with the zoologist Mark Carwardine. Now it's my turn.

00:15 COMM Stephen Fry
Mark and I are heading off to find out exactly what happened to those species that he'd seen dangling on the edge of extinction two decades ago.

Table 16. *Variety of registers - En-Doc 2*

Despite being planned, the language on these instances often contains less formal features, as can also be seen in table 16. These fragments can be re-voiced using voice-over or off-screen dubbing, depending on the market or client.

(3) *Expert interviewee:* interviewees usually appear on-screen and are normally voiced-over in the translated audiovisual product. They do not normally speak from a written text but reply to the questions posed by the interviewer, bearing in mind that they are addressing a wider audience. This means that the language used is spontaneous or semi-spontaneous. As Matamala (2009: 115) points out, this implies that standard language is generally used, containing some informal

features -typical from oral discourse- such as hesitations, false starts, repetitions or anacolutha, i.e. syntactical inconsistencies in a sentence.

(4) *Spontaneous dialogue*: it is normally voiced-over in the Spanish product. It varies in its degree of informality depending on the communicative situation and the speaker's idiosyncrasies: from less informal utterances by a speaker talking to the camera, as if addressing the audience, to more informal dialogue exchanges between participants who are almost unaware that the camera is there. As stated by Matamala (2009: 115), interaction between two people who know each other and who do not directly address the audience are more prone to contain informal language and recurrent hesitations, false starts, repetitions, anacolutha, unfinished sentences, interjections and other oral features.

(5) *Foreign interviewee*: non-native speakers might participate in documentaries as experts. When they appear on screen, they can either speak in English or in their own language. If they talk in English, which is a foreign language for them, their speech may contain errors because of lexical and syntactic interferences, and in some cases, borrowed terms from their mother tongue may appear (see table 17).

01:06:11 Alex Saragoza
The *científicos* were the people who implemented his economic policies. These were the people who wrote the legislation for the passage of laws. These were the people who put together the contracts between the Mexican government and foreign companies and so on. They were elitist, some of them were racist, that is they believed in the notion that the biggest problem that Mexico faced was its backward Indian population.

Table 17. *Variety of registers - En-Doc 7*

If they talk in their own language, sometimes a translation into English is provided in the scripts, as can be seen in table 18, where the interviewee talks in Spanish and the English translation is provided in italics:

01:03:25 Jesus Vargas
La revolución es un proceso social que tiene una relación íntima con toda la historia de México del siglo diecinueve.
The revolution is a social process intimately related to the history of 19th century Mexico.

Table 18. *Variety of registers - En-Doc 8*

To guarantee higher quality levels, MT is normally used with texts using one register. The fact that documentaries tend to combine both formal and informal registers, either planned (based on a written script) or spontaneous, proves more demanding for MT. Additionally, specific features such as some repetitions, hesitations and discourse markers may be more difficult to deal with automatically. Still, when translating documentaries from English into Spanish, it is often the case that many of these features (hesitations, repetitions, etc.) disappear in order to reach voice-over isochrony because informative content is prioritized over expressive features (Orero, 2006). As these features are not usually translated and they make MT processing more difficult, an option would be to delete them, either manually or automatically, from the script that will be fed into the MT engine.

6.3. Terminology

A relevant feature of wildlife documentary films is the inclusion of specific terminology, which varies depending on the topic of the documentary and the general approach, from more to less specialised. Thus, while a documentary film may deal with fishing, another may approach diseases in animals or show the beautifulness of forests and all the fauna and flora they contain. Even if dealing with the same general topic, every wildlife subfield has its specific terminology which may coexist in the same documentary with terminology from other fields.

6.4. Errors and Inaccuracies in the Original Script

As pointed out by Franco et al. (2010) and Matamala (2009, 2010), original scripts can contain errors and inaccuracies. Dates, names of places and terminology may be wrong; text may be missing from the written script or may appear in the wrong place. Possible errors and inconsistencies in the scripts would not affect the work produced by MT engines, although they could slow down the post-editing process. However, if scripts were checked before being machine translated, the number of errors and inconsistencies in the MT output could be minimized and translators would not have to deal with them during the post-editing process.

6.5. Linguistic Inconsistencies in the Original Script

According to Franco et al. (2010: 60), it is not uncommon to find an original script with many linguistic mistakes, poor composition and different ways of spelling the same word; a statement that is also proven in the corpus. In the En-Doc corpus, both spelling (*e.g. though* instead of *thought*) and grammar mistakes (*e.g. worlds* instead of *worlds'*; *this* instead of *these*) have been found, as well as punctuation (*e.g.* interrogation or exclamation marks may appear in the middle of a sentence) and capitalization errors (*e.g.* words without a capital letter may appear after a full stop).

It is also worth stressing that sometimes the script presents the sentences cut into neither non-semantic nor grammatical chunks, as they are fit in different rows (see table 19). When this happens, the semantic and grammatical load of the segments is broken and the MT engine performs worst, as the segment can be split in incoherent syntagmas:

00:25 Listen to the
stories each of us
tells you about
ways of obtaining
Unbelievable
food. Then try to spot the fake from this line-up.

Table 19. *Linguistic inconsistencies. En-Doc*

As Daems et al. (2013) explain, errors in the source text affect the efficiency of MT engines and may influence the quality of the target text even after post-editing. Thus, all the previously described mistakes and segmentation problems inevitably have a bearing on the translation produced by MT engines, and ways to overcome these problems need to be found.

7. Challenges based on the Bilingual Corpus Analysis

An automatic evaluation of the translations produced by eight MT engines and a human-based analysis of the errors found in the MT output was considered an adequate way to predict the challenges of using MT to translate documentary scripts. The results of both the analyses are presented next.

7.1. Automatic Evaluation

BLEU and TER measures were produced to evaluate the 50 sentences translated by the 8 selected MT engines (see table 3). These two measures were chosen as they are the more established among MT researchers at present. On the one hand, and according to Papieni et al. (2002), the higher the BLEU score is, the better the MT output. On the other, the lower the TER is, the better the MT output is, as it means that the error rate is low (Snover et al., 2006). Table 19 presents BLEU and TER scores for each engine:

MT engine	BLEU	TER
Google Translate	29.32	39.41
Apertium	14.19	27.26
Lucy MT	21.20	33.48
Bing	26.88	43.41
Promt	23.99	38.22
Reverso	18.39	25.93
Systran	12.15	3.11
Yandex	27.48	33.63

Table 20. *Automatic evaluation scores*

Results presented on table 20 show that the engines could be divided into four groups according to their BLEU scores. The top quartile would be formed by the MT engines with higher scores Google Translate, *Yandex* and *Bing* (BLEUs from 26.88 to 29.32). The second quartile would include *Promt* and *Lucy MT* (BLEUs from 23.99 to 21.20). In the third, there would only be *Reverso* (BLEU of 18.39), and in the bottom quartile, there would be *Apertium* and *Systran*, the engines with the lower scores (from 12.15 to 14.19). However, if this categorization was made according to TER scores, results would be divided in four groups. The top quartile would include *Bing*, Google Translate and *Promt* (38.22 to 43.41), the middle one would have *Yandex*, *Lucy MT*, *Apertium* and *Reverso* (25.93 to 33.63), and the bottom one would only contain *Systran* (3.11).

The highest BLEU score is reached by Google Translate's engine (29.32 points) and the best TER score is attained by *Bing*'s (43.41 points). BLEU scores do not differ much from scores achieved in other experiments that worked with the same language pair, English > Spanish, within the same translation field of AVT (Nakov, 2008; Kohen et al., 2006; Kohen et al., 2007), as their scores also fluctuated between 23.18 and 35.09. Some of these MT engines achieved better BLEU scores than those presented by the SUMAT project (Bywood, 2013) and are only six points below the eTITLE's results (Melero, 2006). Nevertheless and as an example, the best results are still far from, the ones reached in Vilar et al. (2006), where they presented a BLEU score of 48.6 points when they applied customized MT to subtitling (En <> Spa). It should be taken into account, however, that these results are the first available dealing with documentary film translation and are based on free online engines. Engines created specifically for this domain could, of course, yield better results.

7.2. Human Evaluation

Human evaluation results do not exactly correlate with automatic measures but are to some extent similar. Google Translate is the engine that produces fewer errors (69), followed by *Bing* (78) and *Promt* (82). *Yandex* (102) and *Lucy MT* (114) are the next engines with the fewest errors. The three engines that produce more errors are *Apertium* (151), *Systran* (131), and *Reverso* (129). Thus, if engines were grouped according to their number of errors, the group with the highest scores would include exactly the same engines as in the classifications based on TER and BLEU scores.

Engine	Accuracy		Fluency		Total
	Num.	%	Num.	%	
Google	39	56.52	30	43.48	69
Apertium	87	57.61	64	42.38	151
Lucy MT	59	51.75	55	48.38	114
Bing	30	38.46	48	61.54	78
Promt	45	54.88	37	45.12	82
Reverso	69	53.49	60	46.51	129
Systran	64	48.86	67	51.15	131
Yandex	54	52.94	48	47.06	102
TOTAL	447		409		856

Table 21. *Human Evaluation. Accuracy & Fluency*

As seen in table 21, the majority of errors produced by *Bing* and *Systran*'s engines are related to fluency, while all the other engines have more errors that regard to accuracy. The difference between accuracy and fluency errors produced by *Systran*, *Lucy MT* and *Yandex* is minimal (less than three points between them).

To provide a more detailed analysis, 22 subcategories were considered (12 dealing with accuracy errors and 10 dealing with fluency mistakes), as listed in table 4. No mistakes were found concerning 6 categories: date and time, unit conversion, entity, diacritic accents, punctuation, and unpaired quote marks or brackets. On the contrary, 16 categories reported mistakes: (a) terminology, (b) overly literal, (c) false friend, (d) should not have been translated, (e) number, (f) mistranslations: non-specified errors, (g) omission, (h) addition, (i) untranslated, (j) capitalization, (k) morphology, (l) part of speech, (m) agreement, (n) word order, (o) function words, and (p) unintelligible. Before presenting the results in table 21, an example of each category is presented:

a) Terminology

Original sentence: "Okay, so the next dish is monkey faced eel from Port Baker."
Systran's translation: "La autorización, así que el plato siguiente es anguila hecha frente mono del panadero del puerto."
Back translation: "The authorization, so the dish next is eel done in front of monkey from baker of the port."
Human translation: "De acuerdo, el próximo plateo es anguila caramono de Port Baker."

b) Overly literal

Original sentence: "In a small Ugandan fishing village, nestled along the shores of Lake Victoria, crocodiles have recently killed people."
Reverso's translation: "En un pequeño ugandés el pueblo de pesca, recostado a lo largo de las orillas del lago Victoria, cocodrilos recientemente ha matado a la gente."
Back translation: "In a small Ugandan [from Uganda] the fishing village, nestled along the shores of Lake Victoria, crocodiles have recently killed people."
Human translation: "En un pequeño pueblo de pescadores de Uganda enclavado en la orilla del lago Victoria, últimamente los cocodrilos han matado gente."

c) *False friend*

Original sentence: “Oh, right”
Yandex’s translation: “Oh, a la derecha”
Back translation: “Oh, to the right”
Human translation: “Ah, perfecto”

d) *Should not have been translated*

Original sentence: “Okay, so the next dish is monkey faced eel from Port Baker.”
Apertium’s translation: “Okay, así que el plato próximo es monkey anguila afrontada de Panadero de Puerto.”
Back translation: “Okay, so the dish next is monkey eel faced from Baker of Port.”
Human translation: “De acuerdo, el próximo plato es anguila caramono de Port Baker.”

e) *Number*

Original sentence: “My gun won’t fire. My gun won’t fire.”
Yandex’s translation: “Mis armas no de fuego. Mis armas no de fuego.”
Back translation: “My guns not of fire. My guns not of fire.”
Human translation: “La escopeta no dispara. La escopeta no dispara.”

f) *Mistranslations: non-specified errors*

Original sentence: “She quietly leaves the group and lies down on a secluded spot to await her delivery.”
Lucy’s translation: “Silenciosamente deja el grupo y se tumba en una mancha/sitio retirada para esperar a su entrega.”
Back translation: “She quietly leaves the group and lies down on a secluded spot [patch/place] to await her delivery."
Human translation: “Abandona silenciosamente el grupo y se tumba en un lugar apartado para esperar el momento del parto.”

g) Omission

Original sentence: “But he suspected something else was at work as well.”
Bing’s translation: “Pero sospechaba que algo [missing: más] estaba obrando así.”
Back translation: “But he suspected something [missing: else] was at work as well.”
Human translation: “Pero sospechaba que había algo más.”

h) Addition

Original sentence: “This is better with garlic.”
Systran’s translation: “Esto es mejor con el ajo.”
Back translation: “This is better with the garlic.”
Human translation: “Están más buenos con ajo.”

i) Untranslated

Original sentence: “Then a group of killer whales headed towards shore, as if they intended to strand.”
Apertium’s translation: “Entonces un grupo de killer las ballenas encabezadas hacia shore, cuando si pretendieron a strand.”
Back translation: “Then a group of killer the whales headed [meaning “led”] towards shore, when if intended to strand.”
Human translation: “Entonces un grupo de orcas se dirigió hacia la orilla, como si quisieran quedarse varadas.”

j) Capitalization

Original sentence: “He’s dominated the prairie for some years now, and few have dared comfort him face to face.”
Promt’s translation: “Ha dominado la Pradera durante algunos años ahora, y pocos se han atrevido a oponerse a él cara a cara.”

Back translation: “He's dominated the Prairie for some years now, and few have dared comfort him face to face.”

Human translation: “Ya hace algunos años que domina la llanura y pocos se han atrevido a enfrentarse a él cara a cara.”

k) Morphology

Original sentence: “Between the people, the pavement, and the most overprotective laws in the country.”

Lucy’s translation: “Entre la gente, la acera, y las leyes más sobreproteccionistas del país.”

Back translation: “Between the people, the pavement, and the most overprotectionist laws in the country.”

Human translation: “Entre la gente, el pavimento, y estas leyes tan sobreprotectoras del país.”

l) Part of speech

Original sentence: “Her body strength is recovering quickly, and her calf now kicking.”

Google’s translation: “Su fuerza del cuerpo se está recuperando, y su cría ya patadas.”

Back translation: “Her strength of the body is recovering, and her calf already kick [noun].”

Human translation: “Está recuperando las fuerzas rápidamente y la cría ya le da patadas.”

m) Agreement

Original sentence: “It’s surprising crocs would spend so much energy climbing up this cliff.”

Bing’s translation: “Es sorprendentes crocs pasaría tanta energía subiendo este acantilado.”

Back translation: “It's surprising [plural] crocs would spend [singular] so much energy climbing up this cliff.”
Human translation: “Es increíble que los cocodrilos gasten tanta energía subiendo por este acantilado.”

n) Word order

Original sentence: “Her body strength is recovering quickly, and her calf now kicking.”
Systran’s translation: “Su fuerza del cuerpo se recupera rápidamente, y su becerro ahora dando patadas.”
Back translation: "Her strength of body is recovering quickly, and her calf now kicking."
Human translation: “Está recuperando las fuerzas rápidamente y la cría le da patadas.”

o) Function words

Original sentence: “I feel that it’s so important for me to try to get the Toga people understand what we have in our own back yard is something very unique.”
Google’s translation: “Siento que es tan importante para mí tratar de conseguir [que] la gente Toga entienden [que] lo que tenemos en nuestro propio patio trasero es algo muy especial.”
Comment: In Spanish it is to introduce function words that are not used or necessary in English.
Human translation: “Es muy importante que haga entender a los tonganos que lo que tenemos aquí es algo único.”

p) Unintelligible

Original sentence: “If it’s swimming towards you, get it over the entire head and tighten it up.”

> *Systran's translation*: "Si esto nada hacia usted, conseguirlo sobre la cabeza entera y apretarlo encima de."
> *Back translation*: "If this swims towards you, get it [achieve it] over the entire head and tighten it up above."
> *Human translation*: "Si nada hacia vosotros, la metéis por la cabeza y tensáis."

As shown in table 22, the categories with most errors are (m) *agreement* with 186 cases, (f) *mistranslations: other* with 133, and (i) *untranslated* with 86. While the majority of errors in Google Translate and Apertium are *untranslated* and *mistranslations: other*, all the others engines deal mostly with problems regarding *agreement*. The categories following the lead are (b) *overly literal* with 77 errors, (n) *word order* with 72 and (a) *terminology* with 63. In the central part of the table there are the categories (l) *part of speech* with 56 errors, (g) *omission* with 44, (p) *unintelligible* with 42, (o) *function words* with 39 and (h) *addition* with 37. The categories with lower errors are (j) *capitalization* with 13 errors and (c) *false friends* with 5, as well as three categories with a single error: (d) *should not have been translated*, (e) *number* and (k) *morphology*.

Engine	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**J**	**k**	**l**	**M**	**n**	**o**	**p**	**TOTAL**
Google	5	4	0	0	0	15	2	8	5	0	0	7	13	5	3	2	69
Apertium	11	12	0	0	0	14	5	1	44	0	0	13	23	16	4	8	151
Lucy MT	12	9	0	0	0	21	6	1	10	0	1	6	28	10	4	6	114
Bing	7	1	0	0	0	10	8	1	3	1	0	6	25	5	7	4	78
Promt	7	13	2	0	0	11	3	2	7	2	0	1	18	6	4	6	82
Reverso	7	16	3	0	0	21	9	8	5	5	0	3	28	9	5	10	129
Systran	9	11	0	0	0	25	7	6	6	4	0	3	32	14	9	5	131
Yandex	5	11	0	1	1	16	4	10	6	1	0	17	19	7	3	1	102
TOTAL	63	77	5	1	1	133	44	37	86	13	1	56	186	72	39	42	859

Table 22. *Human evaluation. Types of errors*

To sum up, human evaluation results give us an indication of the most frequent type of mistakes audiovisual translators would have to correct in a post-editing phase: agreement, mistranslated, and untranslated words. Additionally, it indicates that, from the freely available online engines in the English > Spanish combination, Google Translate appears to be the best MT engine, followed by

Bing and *Promt* at least for this study's sample excerpts from documentaries. Although this data may not be relevant for a company deciding to develop their own MT system, (as the analysis is only based on 50 segments and companies normally rely on internal systems specifically developed to satisfy their needs) it is a first step in an underexplored area that might be useful for other scenarios, such as journalistic translation, in which online software can be used.

8. Discussion: Possible Solutions

The bibliographical review and the three corpora analysis have shown several challenges that would have to be addressed in order to integrate MT into the translation process of wildlife documentary films. Before presenting a new workflow to help overcome the challenges, some solutions are proposed for each of the above mentioned challenges.

First of all, solutions regarding the challenges encountered in the bibliographical review –spotting, synchronization and access to the audiovisual content– will be presented. In professional practice, audiovisual translators usually synchronize the visuals and their translation, and are sometimes required to do the spotting, *i.e.* to include the time codes. If MT was to be included in the process of translating documentaries, the MT output would not only have to be corrected during the post-editing stage, but also revised to comply with the various types of synchronies at stake. Correct time codes would also have to be included during the post-editing. In order to do so, full access to the visual content would be required. A suggested scenario to solve these issues would be to include a pre-editing phase (Volk, 2009; Gerlach et al., 2013) in which a time-coded script to be used by translators working into different languages would be created, and additionally, it would be necessary that PE software includes a video player. A tool to limit the maximum number of characters or words per sentence could be also helpful, like PET does for subtitling, as it could help post-editors know how much space they have for each voice-over or off-screen dubbed unit.

Secondly, solutions to the issues found in the analysis of the corpora are proposed. According to the analysis, there are many types of script layouts in English, and to a lesser extent, in Spanish. Therefore, standardizing the script layouts in the original language seems a field in which further work needs to be done. In the meanwhile, creating an MT friendly template every time a documentary is to be translated seems to be a possible solution. This template would contain plain text (not tables) and would be created, again, in a pre-editing phase, ideally with automatic tools that extract the original dialogue from the audiovisual product. It remains to be seen whether this proposed scenario would be feasible when the original documentary is to be translated into one single language or would rather be used in multilingual contexts. Researching this aspect, though, is beyond the scope of this paper.

As for the mixing of various language registers in the same audiovisual programme, a possible solution could be to create a domain-specific engine with wildlife documentaries. Although register-related problems would persist, terminological and lexical problems would hypothetically decrease and reduce the post-editors workload. In order to minimize register challenges, features such as hesitations or repetitions could be erased from the scripts in the pre-editing phase before feeding them into this domain-specific engine.

As for linguistic inconsistencies and errors, they could be rectified either in pre- or post-editing. On the one hand, spelling mistakes and other linguistic problems due to original text formatting could be pre-edited, as they might influence the quality of the MT output. On the other hand, capitalizations and other types of linguistic inconsistencies and errors could be solved during post-editing, as they do not have an impact on the output. Nevertheless, correcting them in the pre-editing phase would be better, as the MT output would drag almost no errors from the original script. In this way post-editors could focus mainly on correcting linguistic errors produced by the MT engine (mainly agreement mistakes and mistranslation, according to our analysis) and solving problems regarding domain-specific issues.

All in all, the analysis has shown that there are problems broadly found in MT which are generally solved through post-editing, but there are also specific challenges related to this text type and audiovisual modality which may be better dealt with in an additional pre-editing phase. What remains to be seen is the impact of this phase in the whole process in terms of time and productivity. However, the availability of a script specifically prepared for MT would have two clear implications. On the one hand, the same script could be used when translating into a different language. On the other, it could let post-editors concentrate more on voice-over and off-screen dubbing specific features. Thus, the following workflow, divided in three steps, is proposed in table 23:

Phase	Tasks
Before translating	1. Build a domain-specific MT engine for wildlife documentary scripts
Pre-editing	1. Spotting 2. Creation of an MT-friendly template 3. Elimination of linguistic inaccuracies 4. Elimination of specific features such as hesitations, repetitions and fluffs.
Machine Translating	1. Machine translate the template
Post-editing	1. Check synchronization between text, images and sound 2. Check register 3. Check terminology 4. Check grammatical and syntactical errors and inaccuracies 5. Solve linguistic inconsistencies especially in terms of accuracy and fluency In order to do so more efficiently, a PE tool including a video display and tool to count words should be used.

Table 23. *Possible solutions. Workflow*

9. Conclusions and Further Research

In conclusion, this article has presented the results of a corpus analysis which has allowed us to identify the main challenges that using MT for the translation of wildlife documentaries might pose: spotting, synchronization, access to audiovisual content, variety on the script format, register variety within a same script, terminology, errors and inaccuracies in the original script, linguistic inconsistencies in the original script, and typical errors in the machine translated output. Three solutions have been proposed to increase the efficiency of post-editing machine translated wildlife documentaries: firstly, pre-editing, as it has been considered to be the answer to challenges such as the inclusion of time-codes, the elimination of certain problematic features (repetitions, hesitations, etc.), and the revision of language of content-related mistakes. Pre-editing has been proposed as a potential solution as it would allow for faster post-editing, an aspect already proven in other contexts such as user-generated content translation (Sertan et al., 2014). Secondly, building a domain-specific engine has been proposed as a possible solution to deal with specific terminology, and thirdly, working with templates has been considered a possible strategy when dealing with a large variety of script formats. Furthermore, the analysis has pointed out the relevance of having access to the audiovisual material, as without it, no successful spotting or synchronization could be made. However, the lack of PE software that allows the inclusion of audiovisual content is still a technical challenge to be overcome. Were all these proposed solutions implemented, post-editing would probably be more efficient and would allow translators to focus on the most specific aspect of this translation mode: synchronisation. Therefore, taking into account the specificities of the genre and the layout characteristics of the scripts, a combination of pre- and post-editing seems to be the most feasible scenario if MT is included in the process of translating wildlife documentary films. Still, further research to prove this hypothesis and its impact on the final workflow needs to be carried out.

Additionally, the analysis has considered a scenario in which a specific engine cannot be built and free online software is used. The analysis of a corpus of machine translated wildlife documentary excerpts has allowed us to identify the main mistakes produced by free online MT engines, namely agreement, mistranslated and untranslated words. This analysis has also shown that, even when using non-specific MT engines, the results of the automatic quality measures are similar to those achieved in other relevant experiments with the same language pair. Such results seem to indicate that future research can be promising as there is still much room for improvement by using, for instance, domain specific MT. Moreover, as many mistakes found in the analysis are of a repetitive nature, and the use of automatic systems to constrain propagation could speed-up the PE task.

To sum up, both the results of the analysis and the presented challenges and solutions seem to indicate that further research on the inclusion of MT in the process of translating wildlife documentaries is advisable. Future investigations could include a similar analysis with other language pairs and translation engines, as well as an analysis of the post-editing effort compared to the human translation effort in which both objective measures and subjective data could be obtained. This future study could also consider other variables such as the inclusion or non-inclusion of a pre-editing phase. All in all, the MT of wildlife documentaries is a novel topic, which opens new research opportunities to which I have tried to contribute by carrying out this exploratory research.

Bibliographical References

Almeida, Giselle de and Sharon O'Brien. "Analysing Post-Editing Performance: Correlations with Years of Translation Experience." *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. Ed. EAMT. (May 2010): 27-28. Available at: http://www.mt-archive.info /EAMT2010-Almeida.pdf [Consulted: 29th June 2016].

Armstrong, Stephen *et al.* "Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation." *Translating and the Computer*. Ed. ASLIB28 (November 2006): 13 pp. Available at: http://www.mt-archive.info/Aslib-2006-Armstrong.pdf [Consulted: 29 th June 2016].

Artegiani, Irene and Dionysios Kapsaskis. "Template files: Asset or Anathema? A Qualitative Analysis of the Subtitles of *The Sopranos*." *Perspectives* 22.3 (2014): 419-436.

Aziz, Wilker *et al.* “PET: a Tool for Post-Editing and Assessing Machine Translation.” *The 8th International Conference on Language Resources and Evaluation (LREC2012)*. Ed. LREC’12 (May 2012): 3982-3987. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf [Consulted: 29th June 2016].

Baldry, Anthony and Paul J. Thibault. *Multimodal Transcription and Text Analysis. A Multimedia Toolkit and Coursebook*. London and Oakville, UK: Equinox, 2006. 270.

Bywood, Lindsay *et al.* “Parallel Subtitle Corpora and their Applications in Machine Translation and Translatology.” *Perspectives: Studies in Translatology. Special Issue: Corpus Linguistics and Audiovisual Translation: in Search of a Integrated Approach* 21. 4 (2013): 595-610.

Daems, Joke *et al.* “On the Origin of Errors: A Fine-Grained Analysis of MT and PE Errors and their Relationship.” *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC2014)*. Ed. LREC’14 (May 2014): 62-66. Available at: http://www.lrec-conf.org/proceedings/ lrec 2014/pdf/532_Paper.pdf [Consulted: 29^{-}June2016].

Díaz Cintas, Jorge and Pilar Orero. “Screen Translation, Voice-over.” *Encyclopedia of Languages*. Ed. Keith Brown. London: Elsevier, 2005. 473-475.

Díaz Cintas, Jorge and Aline Remael. *Audiovisual Translation: Subtitling*. Manchester: St. Jerome, 2007.

Van Dijk, Teun A. “A Note on Linguistic Macro-Structures.” *Linguistische Perspektiven.* Ed. Abraham P. Ten Gate and Peter Jordens. Tübingen: Niemeyer, 1973. 75-87.

Van Dijk, Teun A. and Walter Kintsch. *Strategies of Discourse Comprehension*. New York: Academic Press, 1983.

Eckersley, Helen. “Systems for Evaluating Translation Quality.” *Multilingual Computing and Technology* 3.3 (2002): 39-42.

Espasa, Eva. “Myths about Documentary Translation.” *Topics in Audiovisual Translation*. Ed. Pilar Orero. Amsterdam: John Benjamins, 2004. 183-197.

Fernández-Torné, Anna *et al.* Enhancing Sensorial and Linguistic Accessiblity with Technology: Further Developments in the TECNACC and ALST projects. Paper presented at the *VI Media4All Conference*. (Dubrovnik. 25-27 September 2013).

Fishel, Mark *et al.* "From Subtitles to Parallel Corpora." *Proceedings of the 16th Annual Conference of the European Association for Machine Translation EAMT 2012*. Ed: Cettolo, Mauro *et al.* (May 2013): 3-6. Available at: http://hnk.ffzg.hr/bibl/eamt2012/EAMT-2012.pdf [Consulted: 29-June-2016].

Font Llitjós, Ariadna *et al*. "Error Analysis of Two Types of Grammar for the Purpose of Automatic Rule Refinement." *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*. Ed. AMTA (2004a): 187-196. Available at: http://works.bepress.com/ jaim e_ca rbonell/10/ [Consulted: 29th June 2016].

Font Llitjós, Ariadna and Jaime G. Carbonell. "The Translation Correction Tool: English-Spanish User Studies." *Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC2014).* LREC'04 (2004b): 4 pp. Available at: http://repository.cmu.edu/cgi/viewcontent.cgi? article=1430and cont ext=isr [Consulted: 29th June 2016].

Franco, Eliana P.C. "Documentary Film Translation: A Specific Practice?" *Translation in Context: Selected Contributions from the EST Congress, Granada 1998*. Ed. Andrew Chesterman *et al*. Amsterdam: John Benjamins 2000. 233-242.

—. "Inevitable Exoticism: The Translation of Culture-Specific Items in Documentaries." *La traducción en los medios audiovisuales*. Ed. Frederic Chaume and Rosa Agost. Estudios sobre la Traducció, vol. 7. Castelló de la Plana: Publicacions de la Universitat Jaume I, 2001a. 177-181.

—. "Voiced-over Television Documentaries: Terminological and Conceptual Issues for their Research." *Target* 13.2 (2001b): 289-304.

Franco, Eliana *et al*. *Voice-over Translation: an Overview*. Bern: Peter Lang, 2010.

Freitag, Markus *et al.* "EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project." *Proceedings of the 10th International Workshop for Spoken Language Translation (IWSLT 2013).* IWSLT'13 (2013): 8 pp. Available at:

http://workshop2013.iwslt.org/downloads/EUBRIDGE_MT_Text_Translation_of_Talks_in_the_EUBRIDGE_Project.pdf [Consulted: 29th June 2016].

García Luque, Francisca. "De cómo 'domesticar' un documental de divulgación científica en el proceso de traducción. Estudio de la versión en español de *L'Odyssée de l'espèce*." *SENDEBAR* 22 (2011): 235-263.

Georgakopoulou, Yota. "Challenges for the Audiovisual Industry in the Digital Age: Accessibility and Multilingualism." *MetaNet Forum*. Ed. MetaNet (2010): 1 pp. Available at: http://www.meta-net.eu/events/meta-forum-2010/slides/META-FORUM2010_Georgakopoulou.pdf [Consulted: 29th June 2016].

Gerlach, Johanna *et al.* "Combining Pre-editing and Post-editing to Improve SMT of User-Generated Content." *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*.Eds. EAMT (September 2013): 45-53 Available at: http://www.mt-archive.info/10/MTS-2013-W2-Gerlach.pdf [Consulted: 29th June 2016].

Hurtado Albir, Amparo. *Traducción y traductología*. Madrid: Cátedra, 2001.

Hurtado Albir, Amparo. "Compentence-Based Curriculum Design for Training Translators." *The Interpreter and Translator Trainer* 1.2. (2007): 163-195.

Hutchins, William John. *Machine Translation: Past, Present, Future*. Ellis Horwood Series in Computers and their Applications. Chichester, UK: Ellis Horwood, 1986.

Joscelyne, Andrew. *LSPs in the MT Loop: Current Practices, Future Requirements*. TAUS Report. 2009. Available at: <https://www.taus.net/think-tank/reports/ translate-reports/lsps-in-the-mt-loop-current-practices-future-requirements [Consulted: 29 June-2016]

Kapsaskis, Dionysios. "Professional Identity and Training of Translators in the Context of Globalisation: the Example of Subtitling." *The Journal of Specialised Translation* 16 (2011): 162-184.

Kohen, Philipp and Christov Monz. "Manual and Automatic Evaluation of Machine Translation between European Languages." *Proceedings of the Workshop on Statistical Machine Translation*. Ed. ACL (2006): 102-212. Available at: http://homepages.inf.ed.ac.uk/pkoehn/publications/shared-task-wmt2006.pdf [Consulted: 29-June-2016].

Kohen, Philipp and Hieu Hoang. "Factored Translation Models." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Ed.: ACL (2007): 868-876. Available at: http://homepages.inf.ed.ac.uk/pkoehn/publications/emnlp2007-factored.pdf [Consulted: 29th June 2016].

Koponen, Maarit. "Correctness of Machine Translation: a Machine Translation Post-editing Task." Paper presented at the *3rd MOLTO Project Meeting*. (Helsinki, 2011). Available at: http://www.molto-project.eu/sites/default/files/molto_20110902_mkoponen.pdf [Consulted: 29th June 2016].

Läubli, Samuel *et al.* "Combining Statistical Machine Translation and Translation Memories with Domain Adaptation." *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Ed: NEALT. Proceedings Series, vol. 16. Oslo: Oslo University, 2013. 9 pp. Available at: http://stp.lingfil.uu.se/nodalida/2013/pdf/NODALIDA30.pdf [Consulted: 29th June 2016].

León, Bienvenido. "Science Popularisation through Television Documentary: A Study of the Work of British Wildlife Filmmaker David Attenborough." *5th International Conference of Science and Technology*. Ed. Nigel Sanitt. Berlin: The Pantaneo Forum, 1998: 17-19.

Lommel, Arle. *Multidimensional Quality Metrics*. Paper presented at the *META-FORUM 2013*. (Berlin 2013). Available at: http://www.meta-net.eu/events/meta-forum-2013/talks/arlelommel.pdf [Consulted: 29th June 2016].

Matamala, Anna. *La traducción para voice-over: Online Module for the Master's Degree in Audiovisual Translation*. Barcelona: Universitat Autònoma de Barcelona, 2002.

—. "Teaching Voice-over Translation." *Languages and the Media: New Markets, New Tools. Conference Proceedings*. Ed. ICWE. Berlin: ICWE, 2004. 24-26.

—. "Teaching Voice-over Translation: A Practical Approach." *The Didactics of Audiovisual Translation*. Ed. Jorge Díaz Cintas. Amsterdam: John Benjamins, 2008. 231-262.

—. "Main Challenges in the Translation of Documentaries." *New Trends in Audiovisual Translation*. Ed. Jorge Díaz Cintas. Bristol: Multilingual Matters, 2009a. 109-120.

—. "Translating Documentaries: from Neanderthals to the *Supernanny*." *Perspectives: Studies in Translatology* 17.2 (2009b): 93-107.

—. "Terminological Challenges in the Translation of Science Documentaries: a Case-Study." *Across Languages and Cultures* 11.2 (2010): 255-272.

Matamala, Anna and Carla Ortiz-Boix. *Accessibility and Multilingalism: An Exploratory Study on the Machine Translation of Audio Descriptions*. Forthcoming.

Melero, Melero *et al.* "Automatic Multilingual Subtitling in the eTITLE Project." *Proceedings of ASLIB Translating and the Computer 28*. Ed: ASLIB. (2006). 18 pp. Available at: http://citeseerx.ist.psu.edu/viewdoc/download ?doi=10.1 .1.107.6011andrep=rep1andtype=pdf [Consulted: 29th June 2016].

Nakov, Preslav. "Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Parapharsing, Tokenization and Recasing." *Proceedings of the Third Workshop on Statistical Machine Translation*. Ed. ACL (2008): 147-150. Available at: http://dl.acm.org/citation.cfm?id=1626414 [Consulted: 29th June 2016].

Nord, Christiane. "Text Analysis in Translator Training." *Teaching Translation and Interpreting: Training, Talent, and Experience*. Ed. Cay Dollerup and Anne Loddegaard. Amsterdam: John Benjamins, 1991. 39-47.

—. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St Jerome Publishing, 2014. 154.

O'Brien, Sharon. "Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output." *Across Languages and Cultures* 7.1 (2006): 1-21.

Orero, Pilar. "The Pretended Easiness of Voice-over Translation of TV." *The Journal of Specialized Translation* 2 (2004): 76-96.

—. "Synchronization in Voice-Over." *Aspects of Translation*. Ed. José María Bravo. Valladolid: Universidad de Valladolid, 2006. 255-264.

—. "Voice-over: A Case of Hyper-Reality." *EU High Level Scientific Conference Series. MUTRA Proceedings.*. Ed. MuTra, 2007. 9 pp. Available at:

http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Orero_Pilar.pdf [Consulted: 29th June 2016].

Ortiz-Boix, Carla. *Tecnologies per a l'audiodescripció: estudi sobre l'aplicació de la traducció automàtica i la síntesi de parla a l'audiodescripció en castellà*. Master Thesis. Universitat Autònoma de Barcelona, 2012.

Ortiz-Martínez, Daniel *et al.* "The CASMACAT Project: The Next Generation Translator's Workbench." *Proceedings of the 7th* Jornadas en tecnologia del Habla *and the 3rd Iberian SLTech Workshop (IberSPEECH)*. Ed. ATVS *et al.* (2012): 326-334. Available at: http://www.casmacat.eu/uploads/Main /iberspeech1.pdf [Consulted: 29th June 2016].

Pönniö, Kaarina. "Voice over, narration et commentaire." *Communication audiovisuelle et transferts linguistiques/Audiovisual Communication and Language Transfer*. Ed. Yves Gambier. Special issue of *Translatio* (FIT Newsletter / Nouvelles de la FIT). Strasbourg: Fédération Internationele des Traducteurs (FIT), 1995. 303-307.

Pozo, Arantza del *et al. SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report*. Ed.: Pozo, Arantza del (2012). Available at: <http://cordis.europa.eu/fp7/ict/language-technologies/docs/sumat-annual-report -2012.pdf [Consulted: 29th June 2016]

Ray, Rebecca, ed. *LISA Best Practice Guides. Implementing Machine Translation*. (2004): 74 pp. Available at: http://www.translationoptimization.com/papers/DillingerLommel_MT_BPG.pdf [Consulted: 29th June 2016].

Remael, Alice. "Audiovisual Translation." *Handbook of Translation Studies*, vol. 1. Ed. Yves Gambier, Yves and Luc van DoorslaerAmsterdam: John Benjamins, 2010. 12-17.

Sánchez, Diana. "Subtitling Methods and Team-Translation." *Topics in Audiovisual Translation*, vol 56. Ed. Pilar Orero. Amsterdam/Philadelphia: John Benjamins, 2004. 9-17.

Sertan, Violeta *et al.* "A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation." *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC2014)*. Ed. LREC'14 (May

2014): 1793-1799. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/676_Paper.pdf [Consulted: 29th June 2016].

Sousa, Sheila de *et al.* "Assessing the Post-editing Effort for Automatic and Semi-automatic Translations of DVD subtitles." *Proceedings of Recent Advances in Natural Language Processing*. Ed. ACL (2011): 97-103. Available at: http://aclweb.org/anthology/R11-1014 [Consulted: 29th June 2016].

Uszkoreit, Hans and Arle Lommel. *Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment*. Ed. QT Launch Pad Project (2013). Available at: http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf [Consulted: 29th June 2016].

Vilar, David *et al.* "Error Analysis of Statistical Machine Translation Output." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06).* Ed. LREC (2006): 697-702. Available at: http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf [Consulted: 29th June 2016].

Volk, Martin. "The Automatic Translation of Film Subtitles. A Machine Translation Success Story?" *Journal for Language Technology and Computational Linguistics* 23.2 (2008): 113-125.

Volk, Martin *et al.* "Machine Translation of TV Subtitles for Large Scale Production." *Proceedings of the Second Joint EM+ICNGL Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC '10)*. Ed. Zhechev, Ventsislav *et al.* (2010): 53-62. Available at: http://www.mt-archive.info/10/JEC-2010-Volk.pdf [Consulted: 29th June 2016].

Waibel, Alex. *EU-Bridge Newsletter. 1st Edition.* (2012): 8. Available at: http://project.eu-bridge.eu/img/Newsletter_1_edition.pdf [Consulted: 29th June 2016].

Williams, Malcolm. "The Application of Argumentation Theory to Translation Quality Assessment." *META* 46.2 (2001): 327-344.

Film References

Planet Earth, Alistair Fothergill. Producer: BBC Natural History, UK 2006.

The Crocodile Hunter, John Staiton. Producer: The Best Picture Show Company, Australia 1997-2004.

Chapter 3. Second Article

3. Second Article

Ortiz-Boix, C. and Matamala, A. (Forthcoming). Post-editing wildlife documentary films: A new possible scenario? *JosTrans*. 26.

ABSTRACT: Several studies have proven that, when machine translation followed by post-editing is used to translate general and specialised texts, there is an increase in the productivity, as the post-editing effort is lower than translating *ex novo*. Although the use of machine translation and post-editing has been investigated in Audiovisual Translation, this has never been researched in non-fictional audiovisual genres in which voice-over and off-screen dubbing are applied. Using an English wildlife documentary film as the source text, and Spanish as the target language, this study intends to research whether post-editing involves more or less effort than translating a documentary. Conclusions on the experiment described in this article, in which 12 Audiovisual Translation MA students took part, seem to indicate that post-editing involves less effort than translating.

KEY WORDS: Audiovisual translation, machine translation, post-editing, voice-over and off-screen dubbing.

1. Introduction

In the last two decades, the use of Machine Translation (MT) followed by post-editing when applied to general and specialised translation has been expanding. Such growth has affected not only the market (TAUS, 2009), but also research on post-editing. However, the market of audiovisual translation has barely been affected. Research studies that intend to include MT and post-editing into the process of translating audiovisual products only started a few years ago thanks to European projects such as eTITLE (Melero et al., 2006) or, more recently, SUMAT (Del Pozo et al., 2013), both focusing on subtitling. The promising results presented by the latter led us to believe that applying MT and post-editing to other audiovisual translation modalities might be feasible and worth researching. This has been precisely the aim of the ALST project (Matamala et al., 2012): to investigate the possible application of MT and post-editing into two oral audiovisual transfer modes, namely audio description and voice-over.

The research presented in this article is part of the aforementioned ALST project (FFI-2012-31024), which is financed by the Spanish “Ministerio de Economía y Competitividad”, and focuses exclusively on wildlife documentary films which are translated by means of voice-over and off-screen dubbing. Voice-over is the revoicing of an audiovisual text in another language in which a translating voice is superimposed on the original voice (Franco et al., 2010). It is frequently used in non-fictional audiovisual genres, especially when speakers appear on-screen, but also in fictional TV programmes in Eastern Europe. On the other hand, off-screen dubbing generally refers to the audiovisual transfer mode used to revoice off-screen narrations in which the original voice is substituted by a target language version (Franco et al., 2010). Wildlife documentary films have been selected because, according to a preliminary study by Ortiz-Boix (forthcoming) on a corpus of documentaries, many elements (such as the promising results of the analysed free online MT engines, and the types of errors these engines produce) seem to indicate that it would be feasible to apply MT to this specific genre. However, testing this new scenario in comparison with existing practices with users is yet to be carried out. This is precisely the aim of the research described in this paper: to compare the effort when post-editing a machine translated wildlife documentary and when translating it. Our hypothesis is that post-editing will require less effort than translating.

The article is structured as follows: Section 2 discusses the theoretical approach taken in this paper. In section 3, the methodology used is explained, describing in detail the experiments carried out in June 2014, as well as the methods used to analyse the data. Section 4 discusses the results, taking into account the different types of efforts analysed (temporal, technical, cognitive), and section 5 presents the conclusions and avenues for further research.

2. Theoretical Approach: Post-editing Effort in Audiovisual Translation

This section defines post-editing and how the effort involved has been measured in previous experiments. It also highlights the specificities of the audiovisual transfer modes under analysis.

Post-editing is the "term used for the correction of MT output by human linguists/editors" (Veale and Way, 1997, cited in O'Brien, 2010:1) and, therefore, "the task of the post-editor is to edit, modify and/or correct pre-translated text" (Allen, 2003:297). Post-editing can basically be carried out on two different levels: minimal or light, and full (Allen, 2003:304-306) and, depending on the level of post-editing used, the required effort will vary.

During the last decade, defining and measuring effort within post-editing research has been in the spotlight, thanks to works carried out by Krings (2001), O'Brien (2004, 2005, and 2006) or Martínez (2003), to name just a few. Krings (2001) led the way by determining how to calculate such effort and setting the standard for the majority of the other works on this topic. According to Krings (2001), post-editing effort can be divided into three types: temporal, technical and cognitive. Temporal effort is understood as the time taken to post-edit a document. Technical effort refers to the number of keystrokes, mouse movements and clicks. And cognitive effort applies to "the extent and type of cognitive processes that must be activated to remedy a deficiency in the MT output" (Krings, 2001:179).

While temporal and technical efforts can be directly observed thanks to keylogging software, as can be seen in Allen (2001), Martínez (2003) or Tatsumi and Roturier (2010), cognitive effort cannot be directly observed. Hence, several methods have been used to observe it: Krings (2001) used Think-Aloud Protocols, although he later realised that verbalising all the movements slowed down the process. O'Brien (2004) observed cognitive effort using Translog, a keylogging software. Although Translog did not permit the direct observation of cognitive effort, it did succeed in measuring the number, location and duration of pauses, which were all considered good indicators of cognitive load (O'Brien, 2006; Shreve et al., 2011). Eye-tracking, a non-intrusive equipment that records eye movements and fixations, is another tool used to measure cognitive effort (O'Brien, 2011). To determine the cognitive load of post-editing effort, processing speed, average fixation time and count are generally taken into account. More recently, Lacruz et al. (2014a; 2014b) have claimed that there are two formulae that correlate well with cognitive effort: average pause ratio (APR) and pause to word ratio (PWR). According to them, a low APR (the least possible amount of time spent pausing) combined with a high PWR (the most possible time spent pausing per word) are associated with high levels of cognitive effort. To allow for a lower level of applied cognitive effort, a combination of high APR and low PWR, would be beneficial. Both data can be obtained using keylogging software.

Although an increasing number of researchers study post-editing effort and compare it to translation to determine which one is more productive (Almeida and O'Brien, 2010; Guerberof Arenas, 2009), only a few have analysed post-editing effort as applied to audiovisual translation (de Sousa et al., 2011; Läubli et al., 2013), and specifically to subtitling. Other investigations linking audiovisual translation with post-editing have mostly focussed on the quality assessment of machine translated or post-edited subtitles (Armstrong et al., 2006; Melero et al., 2006; Volk, 2008; Del Pozo et al., 2013 or Bywood et al., 2013).

In order to apply MT and post-editing into the current audiovisual translation workflow, some specificities linked to the genre (wildlife documentary films) and audiovisual transfer modes under analysis (voice-over and off-screen dubbing) need to be taken into account. Voice-over is, together with off-screen dubbing, a modality generally used to translate non-fictional genres in Western Europe (Franco et al., 2010). Among these non-fictional genres, one can find wildlife documentaries, which form the focus of this research. The main characteristics of documentaries are the presence of both a narrator with a generally planned discourse and experts who tend to use a more spontaneous language (Matamala, 2009). Narrators are usually off-screen and dubbed in the target language version, meaning the original narrator cannot be heard and is substituted by a translating voice, whilst on-screen speakers are voiced-over, meaning the translating voice is heard on top of the original, whose sound is lowered down. In both modalities there are synchronisation requirements: translations must take into account the movements and actions on screen (action and kinetic synchronies), and the length of the utterance (isochrony) (Orero, 2006). As far as working conditions are concerned, translators sometimes work without a script or with a script riddled with errors due to the possible lack of post-production scripts (Franco et al., 2010). All these features may be additional challenges when implementing MT in this specific field, as pointed out in a preliminary study by Ortiz-Boix (forthcoming), which suggested pre-editing, as a necessary step for a more successful implementation of MT. Pre-editing (Pym, 1990) is understood as the revision of the format and content of a text before machine translating it. This allows for a higher quality MT output.

3. The Experiment: Methodological Aspects

As stated above, the aim of this experiment was to compare the effort involved in translating and post-editing wildlife documentaries. Following the theoretical approach in section 2, effort was measured in terms of temporal (seconds spent to perform the task), technical (keyboard and mouse usage) and cognitive features (pauses). It was therefore decided that data would be gathered using keylogging software.

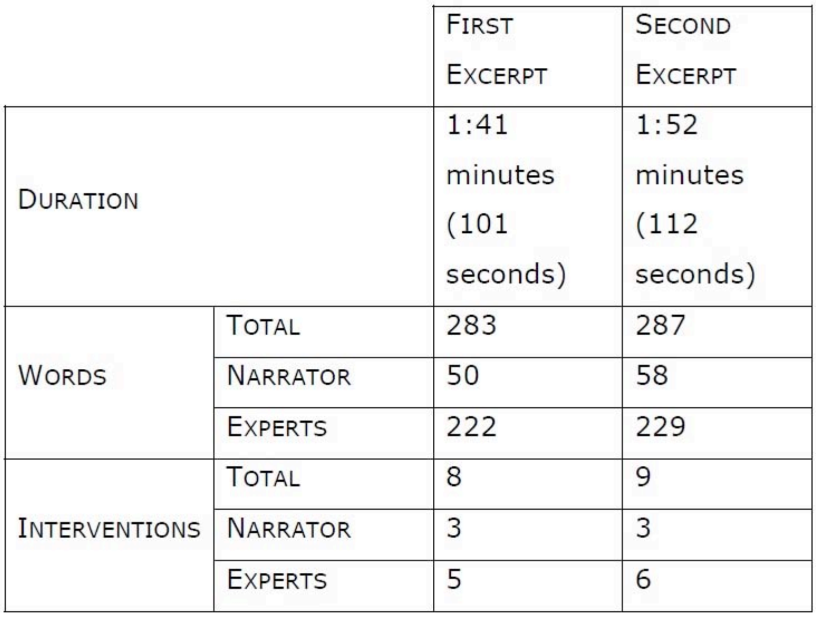
3.1. Participants

12 Master students specialising in audiovisual translation participated in this study. They had all taken a specific course on voice-over, in which they were taught to translate documentaries. Tests were carried out in June, when all participants had successfully finished their courses and were working on their MA thesis. Half of the participants were males and the other half were females, ages ranged between 22 and 27 years old, and all of them had completed a BA in

Translation and Interpreting. They had minimal or no previous experience as professional audiovisual translators and no experience as post-editors. All participants had Spanish as their first language and were highly proficient in English language.

3.2. Materials

Two excerpts of the 7-minute wildlife documentary *Must Watch: A Lioness Adopts a Baby Antelope* were used. They are available on Youtube as an independent documentary (https://www.youtube.com/watch?v=mZw-1BfHFKM) although it is part of the episode *Odd Couples* from the series *Unlikely Animal Friends* by National Geographic (2009). Both excerpts are comparable in terms of length and content, as shown in table 1.



		First Excerpt	Second Excerpt
Duration		1:41 minutes (101 seconds)	1:52 minutes (112 seconds)
Words	Total	283	287
	Narrator	50	58
	Experts	222	229
Interventions	Total	8	9
	Narrator	3	3
	Experts	5	6

Table 1. *Comparison of excerpts*

Both excerpts were machine translated from English into Spanish by Google Translate as, according to previous research by Ortiz-Boix (forthcoming), this is the best free online MT engine to translate wildlife documentary films in this language pair. Automatic measures were calculated with the translations and the post-editings produced by the participants (see Table 2 in 5.3.): BLEU[3]s (Papineni, 2002), h-BLEUs (Snover et al., 2006:224), TERs (Snover et al., 2006) and h-TER[4]s (Snover et al., 2006:224).

3.3. Data gathering Tools

Inputlog (Leijten et al., 2013), a research tool for logging and analysing writing processes developed at the University of Antwerp, was used to record the data. The following measures were obtained: total time, time spent while performing the task and while searching, keylogging, number of mouse movements and clicks, pause thresholds, type of visited internet webpages and type of used software. Although other post-editing tools were considered, they were discarded because they did not integrate audiovisuals (Ortiz-Boix, forthcoming). Inputlog was prioritised over other keylogging software because it allowed for a better simulation of the current workflow of audiovisual translators. It also means that audiovisual materials could be watched without interfering with the tool.

3.4. Test Development

Participants volunteered to take part in the experiment, which was carried out in a lab environment simulating real-life working conditions. They were instructed about the nature of the experiment and signed informed consent forms, following the procedures approved by the Ethical Committee at Universitat Autònoma de Barcelona (UAB). They were instructed that the experiment would develop as follows: they would have to translate an excerpt of a wildlife documentary, and post-edit the machine translated output of another excerpt. They were required

[3] BLEU (Bilingual Evaluation Understudy) and h-BLEU (human targeted Bilingual Evaluation Understudy) are standard automatic measures used to evaluate MT output. The result of these measures arises by comparing MT output with a reference text that can be either its post-editing (BLEU) or a human translation (h-BLEU).

[4] TER (Translation Edit Rate) and h-TER (human targeted Translation Edit Rate) are two other automatic measures used to evaluate MT output. These metrics highlight errors and calculate the edits required in the MT output, in order for the text being edited to resemble a reference text that can be either its post-editing (TER) or a human translation (h-TER).

to use a Microsoft Word template for both tasks, as this was the software used in the MA course they had all taken, but they were free to use any resources available to them online (search engines, video software, etc.). The specific instructions that were given to them were to translate or post-edit, being aware that they had to produce a final document ready to be recorded at a sound studio. They were required to include timecodes in (not out), and they were provided with pre-established timecodes which they could modify if necessary. In the specific case of post-editing, they were instructed to post-edit only when there was a semantic or grammatical error, when some information was omitted or added, and when there were spelling and punctuation mistakes. They were told not to post-edit merely stylistic problems but were asked to rephrase the sentences if, despite being correct, they did not meet the standard conventions of voice-over and off-screen dubbing (this refers to synchronisation features and presentation layouts). After finishing the tasks, they were given a questionnaire on subjective data, the analysis for which is beyond the scope of this paper. Participants were randomly assigned to four different groups in which the two conditions (post-editing/translation) and excerpts (1 and 2) were randomised to avoid any bias regarding the order of presentation.

3.5. Data and Methods

20 valid Inputlog files were collected due to technical problems with four files. Data was obtained from the General Analysis Documents file and exported into Microsoft Excel files. They were analysed using the statistical system R-3.1.2, developed at Bell Laboratories by John Chambers and colleagues.

The following data was obtained for all excerpts and tasks:

(1) *Analysis of temporal effort*: average time spent translating and post-editing, average time spent while working on the Word document, on search engines and using video software.

(2) *Analysis of technical effort*: average number of keyboard and mouse usage, average number of mouse movements and scrolls, average number of mouse clicks and average number of keystrokes. Average number of mouse movements and scrolls, mouse clicks, and keystrokes while working on the Word document, on search engines and on video software were also analysed.

(3) *Analysis of cognitive effort*: average number of pauses and average number of pauses while working on the Word document. To determine PWR and APR, the number of words of each final document and the average time per pause were also assessed.

An ANOVA variance test was used to determine the significance of the results. According to the test, the null-hypothesis can be rejected when the probability value (p-value) is equal or lower than 0.05 ($p<0.05$). The general null-hypothesis of this research states that "there is a significant difference between post-editing effort and translating effort when working with wildlife documentary films scripts."

4. Results

The global analysis indicates that the post-editing effort is significantly lower than the translating effort in the case of technical effort ($F=4.417$, $p=0.050$) and cognitive effort ($F=5.979$, $p=0.025$). However, temporal effort is not ($F=1.297$; $p=0.270$). This may be due to the time one participant spent post-editing, as he spent nearly double the time the others did. When this participant is not taken into account, the post-editing temporal effort is also lower than the translation temporal effort ($F=6.756$, $p=0.019$). Although these results validate our hypothesis, when data from the two different excerpts are analysed in more detail, it can be observed that the difference between post-editing effort and translation effort is not always significant. In the following subsections, and according to the three types of effort identified above, an in-depth analysis is presented.

4.1. Temporal Effort

The analysis of temporal effort indicates that, in the first excerpt, participants spent less time post-editing than translating (see Figure 1): the average time spent translating was 2301.833 seconds (38.36 minutes) and 1853.8 seconds (30.9 minutes) for post-editing. The difference between both tasks being 448.033 seconds (7.47 minutes). ANOVA significance test shows that the temporal effort is significantly lower when post-editing (F=12.940; p=0.006), confirming the results of the general analysis.

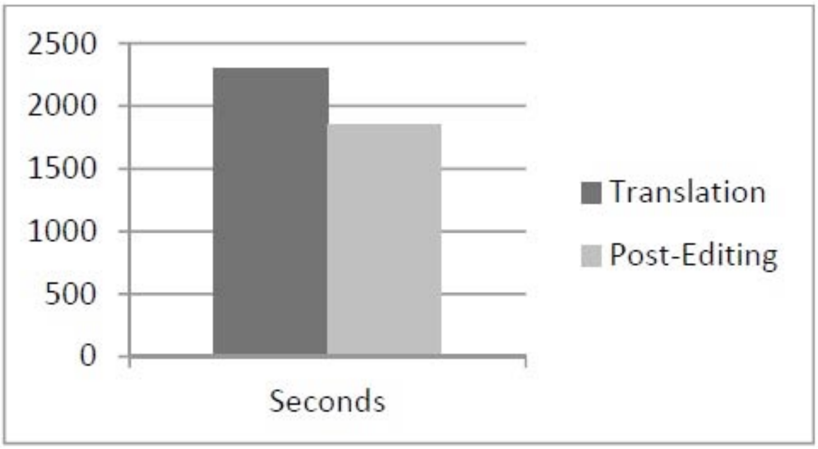


Figure 1. *Comparison of Temporal Effort. Excerpt 1.*

If the timings are explored in more detail, it can be observed (see Figure 2) that, from all the time dedicated to the performance of the translation, participants spent, in excerpt 1, an average of 1556.1438 seconds (25.94 minutes) on the document (67.605% of the time), 477.0633 seconds (7.95 minutes) on search engines (20.725% of the time) and 152.3562 seconds (2.54 minutes) using the video software (6.619% of the time). When post-editing, the difference between the time performing the task on the document (1137.7662 seconds (18.96 minutes), 61.375% of the time) and on the Internet (378.4386 seconds (6.31 minutes), 20.414% of the time) is smaller. Furthermore, post-editors spent more time using video software (165.263 seconds (2.75 minutes), 8.915% of the time). According to the results, there is evidence leading to the belief that post-editors and translators devote approximately the same time to research (F=1.345; p=0.276) and to the video (F=0.034; p=0.612). However, the time spent on each task within the document is significantly different (F=9.918; p=0.012).

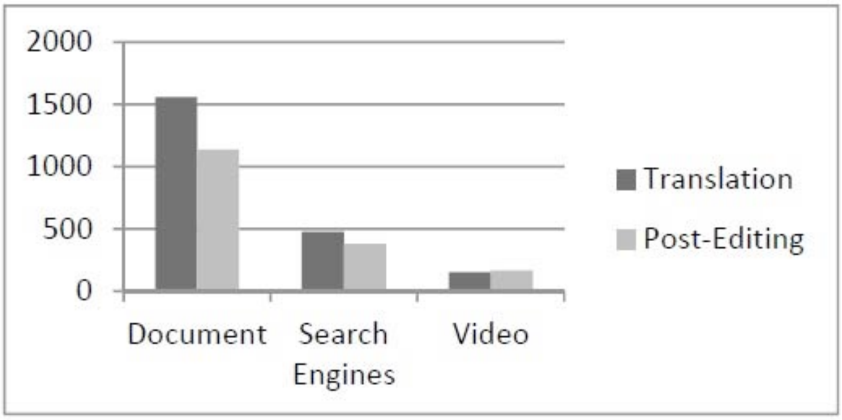


Figure 2. *Division of Temporal Effort. Excerpt 1.*

In the second excerpt, however, the results of the general analysis are not ratified. In this case, the differences between both tasks are minimal (see Figure 3) and the tendency of greater temporal effort when translating does not continue. The average time for translating is 2054.4 seconds (34.24 minutes) and, for post-editing, 2075.25 (34.59 minutes). This means that it took 20.85 more seconds to post-edit this excerpt. Such a change of tendency, as indicated above, is due to the amount of time one of the participants spent post-editing the excerpt. If this participant is considered an outlier and his data is not taken into account for the analysis, the differences are more similar to those of the first excerpt (see Figure 4): 2,054.4 seconds translating (34.24 minutes) and 1,674.6667 seconds post-editing (27.91 minutes), reversing the difference to 379.7333 seconds in favour of post-editing. In this case, ANOVA significance test (F= 0.002; p=0.965) shows that the difference between post-editing and translation in terms of time is not significant. The difference is closer to be significant when the participant who doubled the time is not included in the data (F= 1.265; p=0.304). As this participant's behaviour differed considerably from the others, this participant's results were excluded in the analysis of all the other parameters, which are presented below.

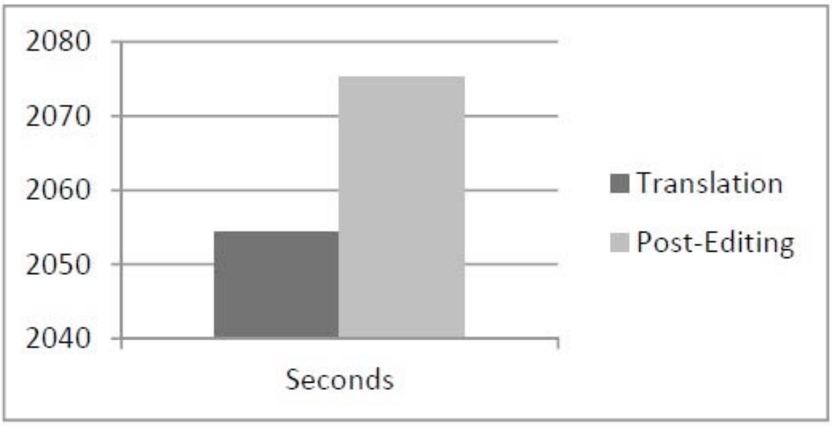


Figure 3. *Comparison of Temporal Effort. Excerpt 2.*

When the temporal effort for the second excerpt is divided into time spent performing the task within the document, on the search engines or on the audiovisual display, the results are slightly different from the ones obtained in excerpt 1 (see Figure 4). Post-editors spent more time working on the document (1357.577 seconds (22.63 minutes), 81.066% of the time) than translators (1222.78696 seconds (20.38 minutes), 59.520% of the time). Post-editors, however, spent less time on the Internet and using the video software (118.9193 seconds (1.98 minutes), 7.101% of the time, and 122.8303 seconds (2.05 minutes), 7.335% of the time, respectively). Translators spent 328.2352 seconds (5.47 minutes, 15.977% of the time) on search engines and 280.6964 seconds (4.68 minutes, 13.663% of the time) on the audiovisual display. The ANOVA significance test shows that there is no significant difference between translation and post-editing in either the Word document (F= 0.355; p=0.573), the search process (F= 3.480; p=0.111) or when working with the audiovisuals (F= 0.562; p=0.482).

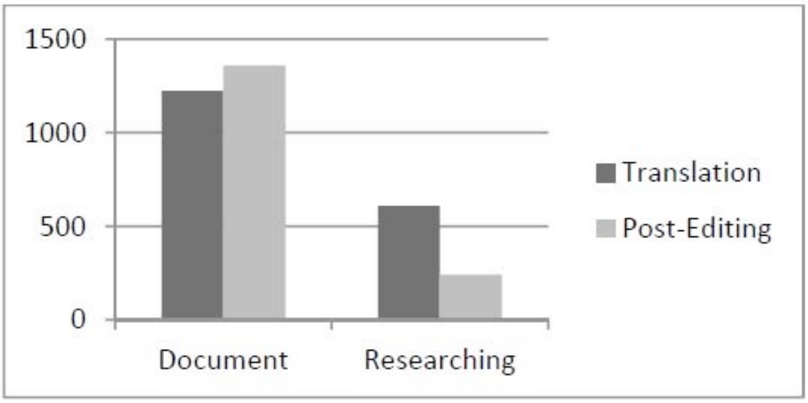


Figure 4. *Division of Temporal Effort. Excerpt 2*

To sum up, although the general analysis indicates that the post-editing temporal effort is lower than the translation temporal effort, a separate analysis of the two excerpts shows inconsistencies. While in the first excerpt the temporal effort is greater in translation than in post-editing, in the second excerpt there are no significant differences between post-editing and translating in terms of temporal effort. In both, no difference can be seen when considering the time spent when performing the task on the document. However, there is also no significant difference in any of the excerpts when considering the time spent both researching and working with the video.

4.2. Technical Effort

The analysis shows that technical effort is higher when translating in both excerpts (see Figures 5 and 6). Translators used the keyboard and the mouse an average of 4079.167 times for the first excerpt and 3972.4 for the second, whilst post-editors used them an average of 2733.8 times for the first excerpt and 2679.333, for the second.

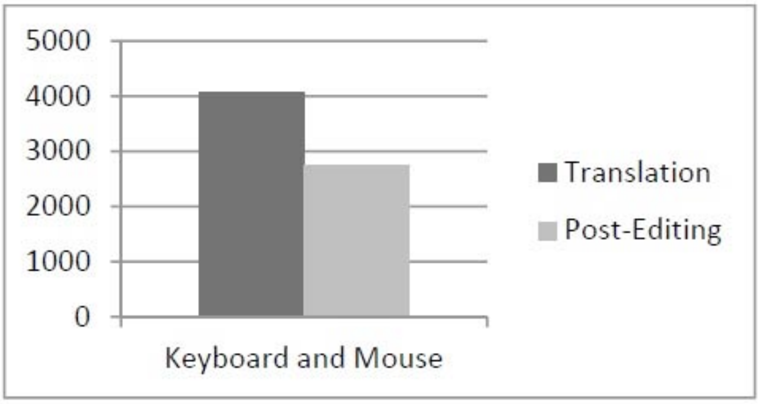


Figure 5. *Comparison of Technical Effort. Excerpt 1*

In the case of the first excerpt, the difference between the use of technical features when translating and post-editing is of 1345.367 keystrokes and mouse movements and clicks (see Figure 5). For the second excerpt, the difference is a little bit lower (see Figure 6): 1293.067.

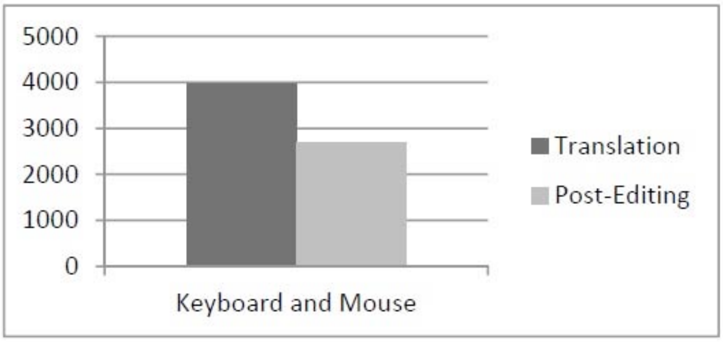


Figure 6. *Comparison of Technical Effort. Excerpt 2*

According to the results there is evidence to suggest that technical effort is higher when translating than when post-editing. However, the difference is only statistically significant in the first excerpt (F=6.365, p= 0.033; excerpt 2: F=3.529, p=0.109). When technical effort is divided into keyboard strokes and mouse usage, these results show that the difference between post-editing and translating technical efforts is due to keyboard use (F=9.943, p=0.012). While the participants who translated the first excerpt used the keyboard an average of 3183 times and the mouse 896.167 times, the ones who post-edited the same excerpt only used the keyboard 1719 times but moved or clicked the mouse more: 1014.8 times (see Figure 7).

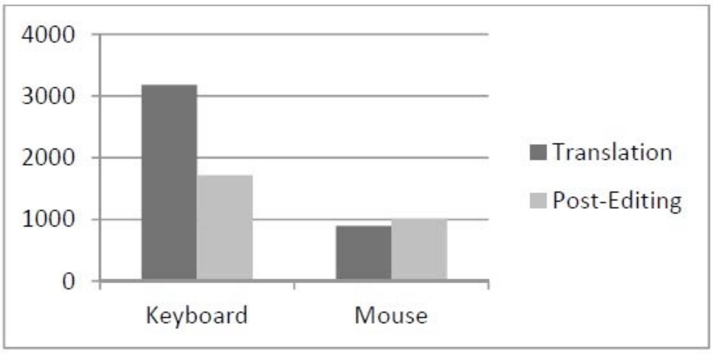


Figure 7. *Division of Technical Effort 1. Excerpt 1*

The tendency to use the mouse more in post-editing is not followed in the second excerpt (Figure 8). Instead, the participants who translated the second excerpt did so. Translators used the keyboard 3029.2 times and the mouse 943.2 times on average; post-editors made an average of 1974.334 keystrokes and 705 mouse clicks or movements (see Figure 8). Despite the translators making 1,000 keystrokes more than the post-editors, the difference in this case is not significant (F= 4.644, p=0.075).

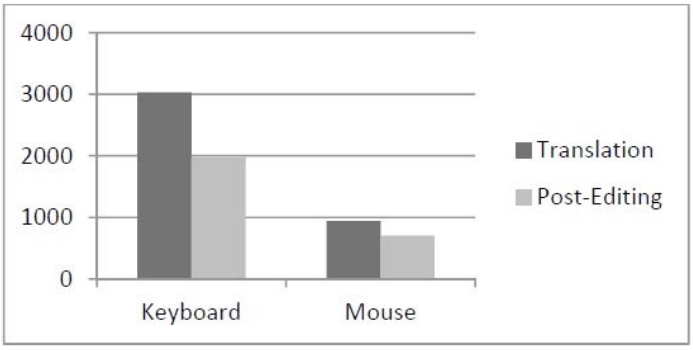


Figure 8. *Division of Technical Effort 1. Excerpt 2*

When analysing the technical effort distribution in the main document, the search engine and the audiovisual display, one can observe that 79.779% of the technical effort (3254.333 keystrokes and mouse movements and clicks) made by the translators of the first excerpt is concentrated on the main document, 17.802% (726.167 keystrokes and mouse movements and clicks) on search engines and only 2.419% of the effort (98.667 keystrokes and mouse movements and clicks) while using the video software. The post-editors who dealt with the same excerpt dedicated almost the same effort to the audiovisual display (3.382%, 92.4 keystrokes and mouse movements and clicks). Their effort on the main document, 4.679 points lower than the translators' (2051.6 keystrokes and mouse movements and clicks), affected the technical effort while searching on the Internet, which reached 21.517% (587.8 keystrokes and mouse movements and clicks). According to these results, it can be stated that a great majority of the technical effort is concentrated in the main document regardless of the task (see Figure 9).

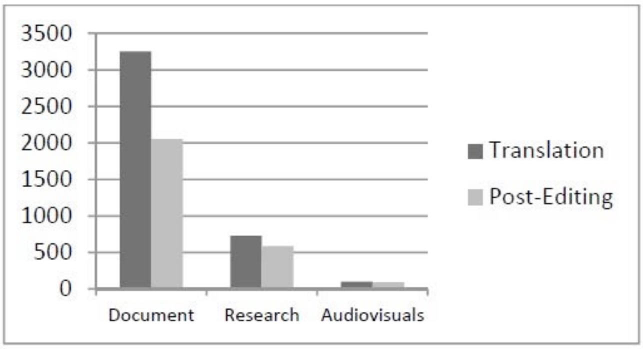


Figure 9. *Division of Technical Effort 2. Excerpt 1*

The results of the second excerpt follow a similar pattern; technical effort is more concentrated in the document and therefore less technical effort is required where research and audiovisual effort is concerned (see Figure 10): when translating, 81.432% of the technical effort (2420.333 keystrokes and mouse movements and clicks) is concentrated in the main document, while 15.935% (214.667 keystrokes and mouse movements and clicks) is dedicated to the search engines and 2.633% (44.333 keystrokes and mouse movements and clicks) to the audiovisual display. In the case of post-editing, 90.333% of the effort (2234.8 keystrokes and mouse movements and clicks) is made on the document, 8.012% (363 keystrokes and mouse movements and clicks) on the Internet and 1.655% (104.6 keystrokes and mouse movements and clicks) while using the video software.

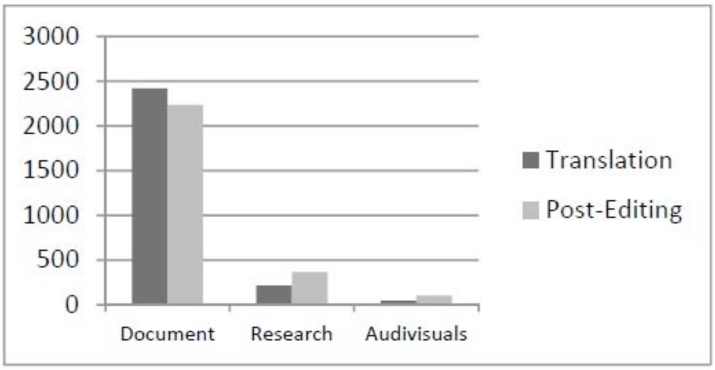


Figure 10. *Division of Technical Effort 2. Excerpt 2*

Apart from showing that technical effort is basically focused on the main document, the in-depth analysis also shows that when translating and post-editing, the use of the keyboard or the mouse varies: keyboard usage is more intensive when working on the document, while it is almost non-existent when working with the video. When doing online searches, the difference between using the keyboard or the mouse is minimal.

When working within the document, the participants who translated the first excerpt (see Figure 11) used the keyboard an average of 2819.334 times (86.633%) and the mouse, 435 times (13.367%). Translators made an average of 355.833 keystrokes (49.002%) and 370.333 mouse movements and clicks (50.998%) while searching on the Internet; and 78.33 keystrokes (7.939%) and 90.833 mouse clicks and movements (92.061%) while using the video software. The ones who post-edited the same excerpt (see Figure 11) made fewer keystrokes (1419 keystrokes, 69.166%) and used the mouse more extensively (632.6 mouse movements and clicks, 30.834%) while working within the document. In the case of using the search engines and the video software, the difference compared with the results of the translators is minimal. They made an average of 294.6 keystrokes (50.119%) and 293.2 mouse movements and clicks (49.881%), and an average of 3.4 keystrokes (3.679%) and 90.833 mouse clicks and movements (92.061%), respectively.

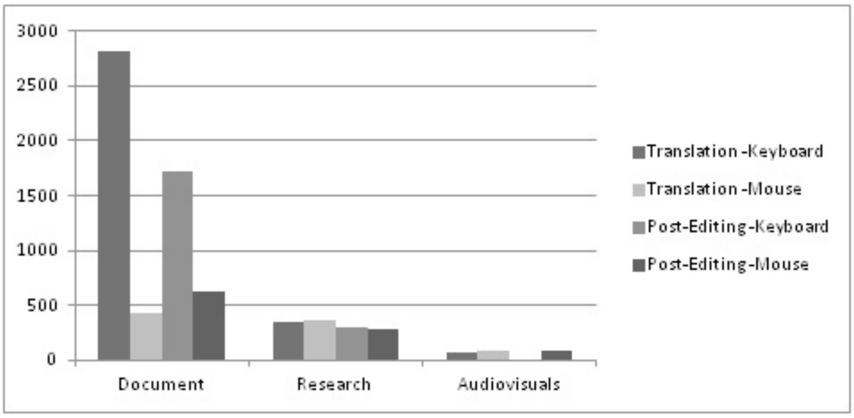


Figure 11. *Division of Technical Effort 3. Excerpt 1*

Regarding the second excerpt (see Figure 12), the results indicate that the trend continues in the case of working within the document and the video software, but the difference between post-editing and translating with regards to technical efforts while searching on the Internet is a bit higher. On the one hand, the translators used the keyboard an average of 2665.8 times (82.410%) and the mouse 569 times (17.590%), when working within the document. In the case of using search engines, they did 361.2 keystrokes (57.062%) and 271.8 mouse movements and clicks (42.938%). Regarding the technical effort while using the audiovisual display, they used the keyboard an average of 2.2 times (2.103%) and the mouse, 102.4 (97.897%). On the other hand, post-editors made 1,855.333 keystrokes (76.656%) and 565 mouse movements and clicks (23.344%) on the document; and used the keyboard 188.667 times (55.279%) and the mouse 96 times (44.721%) on search engines. In the case of the video software, post-editors used the keyboard an average of 0.334 times (0.752%) and the mouse 44 times (99.248%).

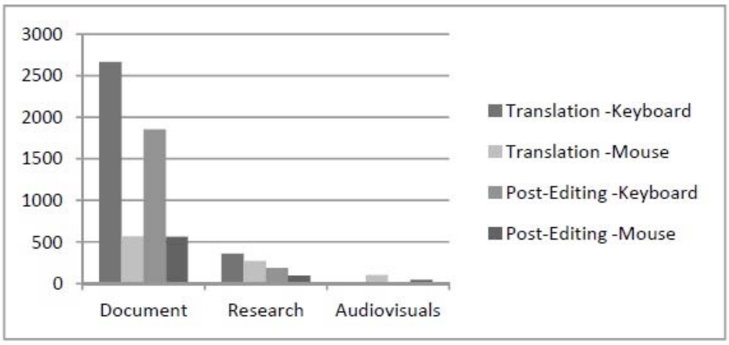


Figure 12. *Division of Technical Effort 3. Excerpt 2*

To summarise, as in the temporal effort, only the first excerpt follows the trend set by the general analysis, which includes both excerpts. The results show that the improvement of the technical effort is due to the decrease in keyboard usage, which is significantly lower only for the first excerpt. Most of the technical effort is concentrated in the main document, where keyboard usage is more intensive.

4.3. Cognitive Effort

Cognitive effort was assessed using the Lacruz et al. (2014a) proposal, which states that the higher the difference between APR and PWR, the more cognitive effort is involved. In order to calculate the APR and the PWR for each task and excerpt, two measures gathered by Inputlog were used: total number of pauses and number of pauses while working on the document.

The results obtained for the first excerpt (see Figure 13) showed that the average APR is 0.191301 in the case of translation and 0.244064 for post-editing. The PWR of the same excerpt is 2.947685 for translation and 1.827491 for post-editing. As discussed in section 2, the lower the APR and the higher the PWR, the more cognitive effort is required during the task. Thus, the bigger the difference between APR and PWR, the greater the cognitive effort. The difference between APR and PWR, aka cognitive effort, is significantly higher when translating[5] (total: 2.756384; only document: 2.123383) than when post-editing (total: 1.583427;

[5] APR and PWR have been calculated using the total number of pauses and with those pauses being made only in the main document. These two conditions have been selected because the first determines the total cognitive effort and the second specifies cognitive effort within the document where technical effort is the focus.

only document: 1.134013) if the total number of pauses are taken into account (F=11.959; p=0.007) or if only the pauses within the document are considered (F=11.332, p=0.008).

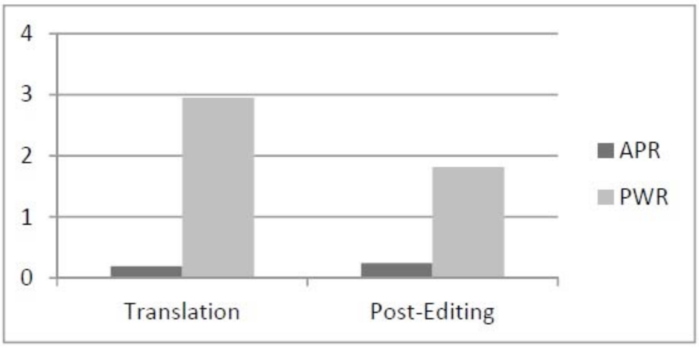


Figure 13. *Comparison of Cognitive Effort. Excerpt 1*

In the case of the second excerpt (see Figure 14), however, the difference between the translation cognitive effort (total: 1.261884; only document: 1.891389) and the post-editing cognitive effort (total: 1.920086; only document: 2.310353) is not significant even when the total number of pauses are taken into account (F=2.712, p=0.151), or when only the pauses while working within the document are chosen (F=4.155, p=0.088).

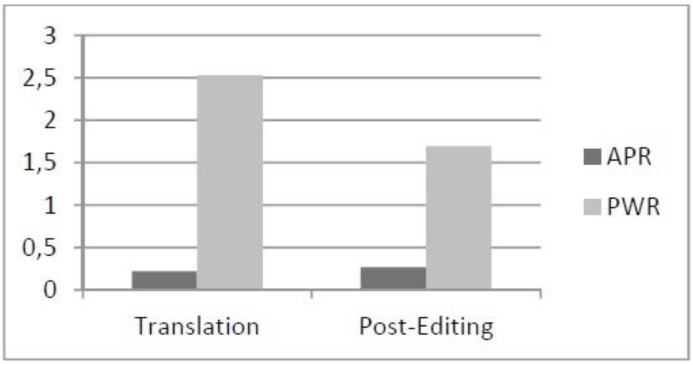


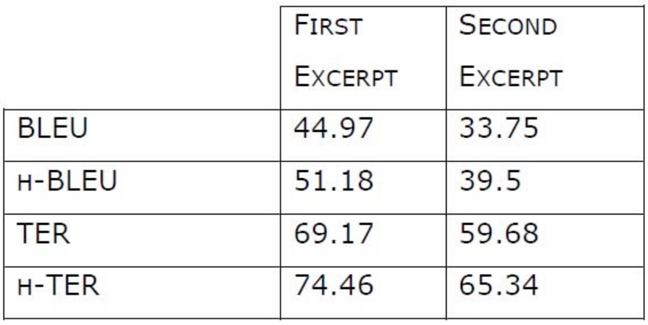
Figure 14. *Comparison of Cognitive Effort. Excerpt 2*

To sum up, the translation cognitive effort is only significant in the case of the first excerpt. However, although the results of the second excerpt are not significant, the translation cognitive effort is also higher.

4.4. Discussion of Results

The results generally confirm the hypothesis that the post-editing effort is lower than the translation effort. Both the general analysis and the analysis of the first excerpt validate the hypothesis, as the temporal, the technical and the cognitive efforts are significantly lower where post-editing is concerned. Nevertheless, the analysis of the second excerpt presents non-significant results. This was unexpected since a previous analysis was carried out to find two comparable excerpts. However, the non-significant results for the second excerpt might be due to three factors:

(1) *Features of chosen documentary*: although comparable in terms of number of words and interventions, the excerpts were not terminologically and syntactically identical. Furthermore, the MT of the second excerpt was worst, as indicated by the BLEU and TER scores presented (see table 2).



	First Excerpt	Second Excerpt
BLEU	44.97	33.75
h-BLEU	51.18	39.5
TER	69.17	59.68
h-TER	74.46	65.34

Table 2. *Automatic Measures*

(2) *Technical skills of the participants*: although all participants had the same training background and were assigned randomly to one of the groups, the analysis shows that the participants who post-edited the second excerpt were probably less skilled with the keyboard than the participants who translated it. This caused an increase in the amount of mouse usage and an increase on the time spent post-editing. Furthermore, the difference was high enough to presume that this

may be the main reason why non-significant differences were observed.

(3) *Amount of data*: the limited number of participants may have had an impact on the significance tests. Therefore, we decided to simulate a situation in which the number of participants who post-edited was hypothetically duplicated. When doubling the number of participants, results are statistically significant only for cognitive effort (F=7.968, p=0.011). Temporal (F=1.249, p=0.296) and technical (F=4.207, p=0.74) efforts, although improving their results in the ANOVA significance test, are still not significant.

5. Conclusions and Further Research

Departing from previous research on post-editing effort, this study built upon the hypothesis that the post-editing effort is lower than the translating effort when working with wildlife documentary films. Global results proved the null-hypothesis of the study. However, results for the second excerpt do not. The excerpt specificities, the uneven technical skills of the participants, and the low number of participants may account for the diverging results.

The data analysis has taken into account the three types of effort specified by Krings (2001), and the following results have been obtained:

(1) *Temporal effort*: the global analysis shows that post-editing is faster. However, results are only statistically significant in the first excerpt.

(2) *Technical effort*: post-editing requires globally less keyboard and mouse usage. Again, the differences are statistically different in the first excerpt but not in the second one.

(3) *Cognitive effort*: post-editing has been proven to be less cognitively demanding although results are not statistically significant in the second excerpt.

Our data also suggests that the effort is concentrated in the main document and it is precisely there where the effort is reduced. In fact, the effort devoted to the search engines or to the audiovisual display does not vary significantly from one task to the other.

In conclusion, the results seem to indicate that it may be possible to use MT followed by post-editing in specific audiovisual genres such as wildlife documentaries which are voiced-over. However, further research should be carried out to confirm the trends shown in this study, which is limited in scope because it only focuses on one language pair (English into Spanish) and has included a small number of participants. Future research could encompass other types of text and include additional language pairs, with their own specificities. It could also take into account other relevant elements such as the subjective opinions and perceived effort of participants. Other aspects worth researching would be the output quality and audience acceptance of post-edited content in comparison with translated products, along with investigations carried out in other translation modalities (Fiederer et al., 2009). It would also be highly relevant to measure the professional performance efforts of audiovisual translators. All in all, there are many aspects to be researched but this article has hopefully been a first step towards future studies on the implementation of translation technologies in the field of audiovisual translation and media accessibility, an area that is still under-researched especially when oral modalities such as voice-over, dubbing or even audio description are concerned.

Bibliography

Allen, Jeff (2001). "Postediting: an integrated part of a translation software program." *Language International*, 13(2), 26-29.

— (2003). "Post-editing." *Benjamins Translation Library*, 35, 297-318.

De Almeida, Gisela and Sharon O'Brien (2010). "Analysing post-editing performance: corrections with years of translation experience." *Proceedings of the 14th annual* *conference of the European association for machine translation, St. Raphaël, France*.

Armstrong, Stephen; Colm Caffrey; Marian Flanagan, Minako O'Hagan; Dorothy Kenny and Andy Way (2006). "Improving the Quality of DVD Subtitles via Example-Based Machine Translation." *Proceedings of the Translating and the Computer 28 Conference*, London, England.

Bywood, Lindsay; Martin Volk; Mark Fisheland Panayota Georgakopoulou (2013). "Parallel subtitle corpora and their applications in machine translation and translatology." *Perspectives* 21(4), 595-610.

Del Pozo, Arantza; Gerard van Loenhout; Anthony Walker; Panayota Georgakopoulou and Thierry Etchegoyhen (2013). *SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report.*

Fiederer, Rebecca and Sharon O'Brien (2009). "Quality and machine translation: A realistic objective." *JoSTrans,The Journal of Specialised Translation*, 11, 52-74.

Franco, Eliana; Anna Matamala and Pilar Orero (2010). *Voice-over translation: An overview*. Peter Lang.

Guerberof Arenas, Ana (2009). "Productivity and quality in MT post-editing." *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*.

Krings, Hans P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*. (Vol. 5). Kent State University Press.

Läbuli, Samuel; Martin Fishel; Martin Volk and Manuela Weibel (2013). "Combining Statistical Machine Translation and Translation Memories with Domain Adaptation." *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, May 22-24, 2013, Oslo University, Norway. NEALT Proceedings Series 16.

Lacruz, Isabel; Michael Denkowski and Alon Lavie (2014a). "Cognitive Demand and Cognitive Effort in Post-Editing." *Third Workshop on Post-Editing Technology and Practice*.

— (2014b). "Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter." *EACL 2014*.

Leijten, Marielle and Luuk Van Waes (2013). "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes." *Written Communication 30*(3), 358–392.

Martínez, Lorena G. (2003). *Human translation versus machine translation and full post-editing of raw machine translation output*. MA Diss. Dublin City University, http://sceuromix.com/enlaces/MASTER%20IN%20TRANSLATION%20STUDIES%20BY%20LORENA%20GUERRA-2003.pdf (consulted 10.05.2016).

Matamala, Anna (2009). "Main Challenges in the Translation of Documentaries." In Jorge Díaz Cintas (ed.). *New Trends in Audiovisual Translation*. Bristol: Multilingual Matters, 109-120.

Matamala, Anna ; Anna Fernández-Torné. and Carla Ortiz-Boix (2012). "Technology and AD: The TECNACC Project." *Languages and the Media 2012*, Berlin. < http://ddd.uab.cat/pub/presentacions/2012/117159/fernandez_matamala_ortiz_berlin2012.pdf > (last accessed: 16th May 2016)

Melero, Maite; Antoni Oliverand Toni Badia (2006). "Automatic multilingual subtitling in the eTITLE project." *Proc. of the 28th International Conference on Translating and the Computer*, 28 16-17 November 2006 in London. London: ASLIB.

O'Brien, Sharon (2004). "Machine translatability and post-editing effort: How do they relate." *Proc. of the 26th International Conference on Translating and the Computer*, 18-19 November 2004 in London. London: ASLIB.

— (2005). "Methodologies for measuring the correlations between post-editing effort and machine translatability." *Machine Translation*, 19(1), 37-58.

— (2006). "Pauses as indicators of cognitive effort in post-editing machine translating output." *Across Languages and Cultures*, 7(1), 1-21.

— (2010). "Introduction to Post-Editing: Who, What, How and Where to Next." *The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

— (2011). "Towards predicting post-editing productivity." *Machine Translation*, 25(3), 197-215.

Orero, Pilar (2006). "Synchronisation in Voice-over." *New Spectrum in Translation Studies* 255-264.

Ortiz-Boix, Carla (Forthcoming). "Post-Editing Wildlife Documentaries: Challenges and Possible Solutions."

Papineni, Kishore; Salim Roukos; Todd Ward and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics (ACL), 311-318.

Pym, Peter J. (1990). "Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system." *Translating and the Computer* 10 (1990), 80-96.

Shreve, Gregory M.; Isabel Lacruz and Erik Angelone (2011). "Sight Translation and Speech Disfluency: Performance Analysis as Window to Cognitive Translation Processes." *Methods and Strategies of Process Research*, Jon Benjamins, 121-146.

Snover, Matthew, Bonnie Dorr; Richard Schwartz; Linnea Micciulla and John Makhoul (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*, 8-12 August 2006 in Cambridge, Massachusetts, USA. Massachusetts: AMTA, 223-231.

De Sousa, Sheila C.; Wilker Aziz and Lucia Specia (2011). "Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles." *RANLP*, 97-103.

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

TAUS (2009). *LSPs in the MT loop: current practices, further requirements*. < https://www.taus.net/think-tank/reports/translate-reports/lsps-in-the-mt-loop-current-practices-future-requirements> (last accessed: 18th May 2016)

Veale, Tony and Andy Way (1997). "Gaijin: A bootstrapping, template-driven approach to example-based MT." *Proceedings of the New Methods in Natural Language Processing (NeMNLP97)*. Sofia, Bulgaria: 239-244.

Volk, Martin (2008). "The Automatic Translation of Film Subtitles. A Machine Translation Success Story?" *Journal for Language Technology and Computational Linguistics*, 23(2), 113-125.

Filmography

National Geographic (ed.) (2009). "Must Watch: A lioness adopts a baby antelope". *Unlikely Animal Friends*. Episode: "Odd Couples."

Acknowledgements

This article is part of the *ALST project* (reference FFI-2012-31024), directed by Anna Matamala and funded by the Spanish "Ministerio de Economía y Competititvidad," and also of research group TransMedia Catalonia (2014SGR27). This article is also part of the research carried out by Carla Ortiz-Boix under the supervision of Dr Anna Matamala within the PhD in Translation and Intercultural Studies (Department of Translation, Interpreting, and East Asian Studies) at Universitat Autònoma de Barcelona. The data analysis was carried out during a research stay at Dublin City University. We would also like to thank Dr Sharon O’Brien for welcoming Carla Ortiz-Boix on a research stay at Dublin City University.

Chapter 4. Third Article

4. Third Article

Ortiz-Boix, C. and Matamala, A. (forthcoming). Assessing the Quality of Post-edited Wildlife Documentaries. *Perspectives. Studies in Translatology*.

ABSTRACT: This article presents the results of an experiment to assess the quality of post-edited wildlife documentary films to be voiced-over and off-screen dubbed, which was compared to the quality of human translation. The main hypothesis of the article is that there are no significant differences between translated and post-edited texts in terms of quality. Twelve MA students translated and post-edited two excerpts of an English wildlife documentary into Spanish. Then, six professional translators assessed both the translations and post-edited texts by: (1) grading the documents, (2) correcting them using a Multidimensional Quality Metrics-based error classification that takes into account documentary translation specificities, and (3) answering questionnaires on their impressions. Results confirm the main hypothesis by indicating that the quality of post-edited and translated wildlife documentary films is significantly similar.

KEY WORDS: Screen Translation; Voice-Over; Translation of Documentaries; Machine Translation; Post-Editing; Quality Assessment

1. Introduction

The use of machine translation (MT) followed by post-editing (PE) has been expanding in the translation industry and has been increasingly researched in the last few decades. Several projects investigating the possible inclusion of MT and PE into the process of translating audiovisual products started almost a decade ago. Such projects (e.g. eTITLE or SUMAT) were funded by the European Commission and focused basically on subtitling. The results obtained, especially by the SUMAT project (Etchegoyhen et al., 2014), and the lack of research on the implementation of MT systems in other audiovisual translation modes inspired the ALST project (Matamala et al., 2012) to investigate the possible application of MT and PE into audio description and voice-over, two audiovisual transfer modes which are delivered orally.

The research presented in this article, which is part of the ALST project, focuses exclusively on wildlife documentary films translated by means of voice-over and off-screen dubbing. Voice-over is the revoicing of an audiovisual text in another language in which a translating voice is heard on top of the original voice (Franco et al., 2010:43). It is frequently used in non-fictional audiovisual translation genres, especially when speakers appear on-screen, but also in fictional TV programs in Eastern Europe. Off-screen dubbing, conversely, refers to the audiovisual transfer mode generally used to revoice off-screen narrations in which the original voice is substituted by a target language version (Franco et al., 2010:41).

This article aims to compare the quality of post-edited texts with the quality of human translations and can be considered the follow-up to the investigation presented in Ortiz-Boix and Matamala (forthcoming), where the effort involved in post-editing a wildlife documentary excerpt was compared to the effort involved in translating it. Results showed that the post-editing effort is less than the human translation effort. However, it remains to be seen whether the output quality produced during a post-editing process is similar to the quality produced in a translation process. This article presents the results of an experiment which aimed to validate the hypothesis that there are no significant differences between the quality of post-edited texts and the quality of translations.

The article is structured in six Sections: Section 2 discusses the theoretical approach of the paper. In Section 3, the methodology used in the experiment is explained, as well as the methods used to analyse the data. Sections 4 and 5 discuss the results and, finally, Section 6 presents the conclusions and proposes further research.

2. Quality Assessment in Translation, Machine Translation and Post-editing

Quality has been a central issue in Translation Studies since the beginning of the discipline and many studies have dealt with it (e. g. Nida, 1964; Carrol, 1966; House, 2006; Koller, 1987; Toury, 1995; Gambier, 1998; Hansen, 2008; Chiaro, 2008). Quality assessment (QA) is product-based and is approached differently depending on the theory of translation that lies behind each QA model (House, 2006). House (2006) divides the different approaches into several categories: (1) anecdotal approaches, which are based on reflections of professional translators who are mainly concerned with the text being faithful to the original (e. g. Savoy 1968); (2) neo-hermeneutic approaches, which consider that the quality of a translation depends on how fully a translator identifies with the original text (e. g. Kupsch-Loseriet, 1994); (3) response-oriented approaches, which are communicatively oriented (e. g. Nida, 196; Nida and Taber, 1969; Carroll, 1966); (4) text- based approaches, which can focus on linguistic (Reiss 1978), literary (Toury, 1985) or functional aspects (Reiss and Vermeer, 1984) of the translation, and (5) pragmatic linguistic approaches, which analyse the linguistic-situational particularities of source and target texts, compare them and assess their relative match to assess quality (House, 2006). Depending on the approach, various methods have been proposed, such as quantitative and qualitative testing by competent judges, comparing translations against reference models, rating sentences according to pre-established scales of intelligibility and informativeness and gathering respondents' opinions, among others.

Studies on MT have also addressed quality and QA as a main issue (see e.g. Hutchins and Somers, 1992; Krings, 2001; King et al., 2003; Fiederer et al., 2009; Armstrong et al., 2006). While the first studies only used human judges to evaluate the MT output, measures to assess the quality automatically by means of a preprogrammed tool (automatic measures) have been expanding. Such measures compare and correlate translations or post-edited texts with the MT output to set the quality of the machine translated texts. Thus, automatic measures still need human translations: on the one hand, there are measures of precision such as BLEU (Papineni et al., 2001), H-BLEU (Snover et al., 2006), NIST (Snover et al., 2008) or METEOR (Lavie et al., 2009), which compare the MT output with reference translations or post-edited texts. On the other hand, there are editing-distance measures such as TER (Snover et al., 2006), H-TER (Snover et al., 2006), WER (Jiménez Linares, 2008) or PER (Jiménez Linares, 2008), which calculate the number of modifications needed on a MT output so that it resembles a reference translation or post-editing.

Apart from automatic measures, human-based evaluations have been carried out on raw MT output (e. g. O'Brien, 2005; Aziz et al., 2012). The majority of these evaluations focus on the fidelity or accuracy of the MT output (e. g. Arnold et al., 1994; Dabbadie et al., 2002; Roturier, 2006; Fiederer, 2009), its intelligibility or clarity (e. g. Hutchins and Somers, 1992; Fiederer et al., 2009), and its style (Hutchins and Somers, 1992; Arnold et al., 1994; Fiederer et al., 2009). Others focus on using post-editing as a measure of assessment (e.g. Popovic et al., 2013) or on classifying the errors produced by the MT engines (e.g. Federman, 2012). Regarding error classification, Lommel et al. (2014) designed the Multidimensional Quality Metrics (MQM), based on functional theories of translation, which propose several error issue types dealing with both the micro and the macrostructure of the text. This metrics can be used to assess MT output, but also post- edited texts and translations (Lommel et al., 2014).

Regarding research on the specific topic of our investigation –evaluating the quality of post-editing in comparison to translations– it is rather limited. Plitt et al. (2010), Guerberof Arenas (2009, 2012), Fiederer et al. (2009), Carl et al. (2011), García (2011) and De Sutter et al. (2012) have compared, to a greater or lesser extent, the quality of post-edited texts and translations.

Plitt et al. (2010) assessed the quality at Autodesk, a company whose Localization Service department actively uses MT and measures the usefulness of MT, which can be used as a translation productivity tool or for gisting. Autodesk’s translation QA team reviewed part of the work of 10 out of 12 participants, who translated and post-edited randomly selected samples of translated and post-edited texts from English into French, Italian, German and Spanish, and rated them in two levels: "average" or "good", depending on whether they would publish the texts as they read them. The results presented in Plitt's article showed that translations contained a higher number of mistakes than post-edited sentences in all four languages.

Guerberof Arenas (2009, 2012) used three reviewers to blindly assess human translated and post-edited MT segments, as well as corrected segments previously extracted from a translation memory, on the topic “business intelligence technology" by using the LISA QA model. These three reviewers measured and classified the number of errors in eight categories: mistranslation, accuracy, terminology, language, style, country, consistency, and format. During the analysis, the number of errors per source (human translated segments, post-edited MT segments or corrected segments extracted from a translation memory) was calculated. Guerberof Arenas’ (2012) results show that the quality produced by translators was significantly higher when they post-edited a segment produced by the MT engine or proposed by the translation memories. It was also observed that while the majority of language, terminology and style errors were found in the segments translated from scratch, the majority of accuracy errors were seen in the corrected segments extracted from translation memories, and mistranslation errors were mainly present in post-edited MT segments.

As part of a wider study, Fiederer et al. (2009) assessed the quality in machine-translated texts by evaluating 30 source sentences with three translated and three post-edited versions according to three parameters: clarity, accuracy and style. The sentences were assessed by 11 raters, who ranked the translations and post-edited texts from 1 to 4 (being 1 the lowest mark and 4 the highest). They were also asked to indicate their favorite translated option out of the six proposals for each source sentence. Evaluators scored translated and post-edited texts equally in terms of clarity. However, post-edited texts were evaluated higher with regards to accuracy, and translations were evaluated higher when style was considered. All in all, raters chose the translated sentences as their favorites.

Carl et al. (2011) presented a study on the post-editing experience of translators working from English into Danish. It included the QA of three general texts (850 characters all together), evaluated by seven native speakers of Danish. Each rater ranked two human translations and two post-edited texts. Post-edited texts were found to be better than translations, although the difference was not statistically significant.

García (2011) explored post-editing in non-professional contexts in the English-Chinese language pair. In order to do so, one rater assessed the quality of a 500-word text both translated and post-edited by using the Australian National Accreditation Authority for Translators and Interpreters' (NAATI) guidelines. The results presented in the study show that post-edited passages were of higher quality than the translated.

Finally, De Sutter et al. (2012) studied the quality of a text half translated half post-edited by 15 translation trainees from English into French. The assessment was done by a single evaluator who rated the segments using a five-point scale. The results show that translations receive higher scores than post-edited texts, although the difference between the translations and post-edited texts was minimal.

Taking into account all this prior work on the topic, a mixed-approach model for QA, based on both text-based and response-oriented models, has been proposed for the current experiment, as described next.

3. Methodology

The aim of the experiment is to assess the quality of post-edited wildlife documentaries compared to the quality of human translations. It is built upon the hypothesis that there is no significant difference between the quality of post-edited and the quality of translated wildlife documentaries to be voiced-over and off-screen dubbed from English into Spanish.

3.1. Participants

The evaluators participating in this study were six lecturers of audiovisual translation MAs in Spanish universities who are experts on voice-over and currently work or have worked as professional voice-over translators. Participants' profiles are comparable, as all of them have a BA in Translation Studies except for one, who has a BA in German Philology. Five of them have either attended PhD courses or have a PhD in Translation. Furthermore, they all work from English into Spanish. Two participants are more experienced than the others: when the experiment was carried out, participants 1 and 5 had worked as audiovisual translators for 16 years and had taught for 11 and 5 respectively. In comparison, participants 2, 4 and 6 had between 5 and 8 years of experience as audiovisual translators and had taught for the last 4 or 5. Participant 3 had worked as audiovisual translator for 10 years and taught for the last 8. The number of raters is limited but in line with previous research in QA of MT output using human judges (e. g. De Sousa et al., 2011) and even higher than existing post-editing experiments (Guerberof Arenas, 2009; Guerberof Arenas, 2012; García, 2011 and De Sutter et al., 2012).

3.2. Materials

The materials used were 6 translations and 6 post-edited texts of two excerpts of the 7-minute wildlife documentary film *Must Watch: a Lioness Adopts a Baby Antelope* (i.e. a total of 24 documents*)*. The translations and post-edited texts were produced by 12 students of an MA in audiovisual translation who had taken a specific course in voice-over where they were taught how to translate documentaries. The documentary is an excerpt from the episode *Odd Couples* from the National Geographic series *Unlikely Animal Friends* broadcast in 2009 and currently available as an independent video on YouTube[6]. The excerpts are similar in length, number of words and entries (Table 1). A short transcription of one of the excerpts used for the experiment is included in the following lines as an example of the type of text used:

[6] https://www.youtube.com/watch?v=mZw-1BfHFKM (Last accessed: 24th February 2016)

> *00:00* ***Narrator:*** *For days the calf wandered looking for its herd, while the lioness followed.*
>
> *00:06* ***Saba:*** *Of course, every oryx it saw was potentially its mother and potentially food and life. So, it would constantly try to rejoin adult oryxes. Well, Kamunyak would allow it to go a certain distance away but as soon as it started to move off with the oryxes she was then up on the warpath.*
>
> *[…]*

Both excerpts were machine translated by Google Translate, as a pre-analysis by Ortiz-Boix (forthcoming) proved this was the best free online MT engine that can be used to translate wildlife documentary scripts at the time the experiment took place. The pre-analysis compared the output produced by Google Translate to the output of 7 other free online MT engines using automatic quality measures and a human analysis of errors. The MT of the first excerpt was slightly better than the MT of the second, according to automatic metrics such as BLEU[7], h-BLEU[2] (Papieni, 2002), TER[3] and h-TER[8] (Snover et al., 2006) (see also Table 1).

	1st Excerpt	2nd Excerpt
DURATION	101 seconds (1:41 minutes)	112 seconds (1:52 minutes)
WORDS	283	287
ENTRIES/LINES OF SPEECH	8	9
BLEU	44.97	33.75
h-BLEU	51.18	39.5
TER	69.17	59.68
h-TER	74.46	65.34

Table 1. *Comparison of excerpts*

[7] BLEU (Bilingual Evaluation Understudy) and h-BLEU (human targeted Bilingual Evaluation Understudy) are standard automatic measures used to evaluate MT output. The result of these measures arises by comparing MT output with a reference text that can be either its post-editing (BLEU) or a human translation (h-BLEU). The given result is between 0 and 100, where the higher the score is, the better the translation is considered.

[8] TER (Translation Edit Rate) and h-TER (human targeted Translation Edit Rate) are two other automatic measures used to evaluate MT output. These metrics highlight errors and calculate the edits required in the MT output, in order for the text being edited to resemble a reference text that can be either its post-editing (TER) or a human translation (h-TER). The given result is between 0 and 100, where the lower the score is, the better the translation is considered.

3.2.1. Test Development

Participants carried out the experiment from their usual place of work. They were given detailed instructions to assess 24 documents in 20 days without knowing which of them were translations or post-edited texts. The experiment was divided in two parts:

(1) For the first part they were given just one day: they were instructed to read each document and grade it according to their first impression on a 7-point scale (scoring round 1). The order of the documents was randomized. It was expected that this approach would reflect more accurately how the audience would react to the documentary and provide interesting findings on the quality of the text from a target audience perspective.

(2) In the second part, they were asked to review and correct the documents, identify the errors following a specific evaluation matrix (see Section 3.4) and grade the documents after the correction on a 7-point scale (scoring round 2). The approach in this second round was more academic, as reviewers were given a set of specifications that guided them through a more didactic analysis. Afterwards, they were instructed to answer an online questionnaire on their opinion about the document they corrected (questionnaire-based evaluation, see Section 3.4.), and gave each document a final grade between 0 and 10 (scoring round 3). These two last grading rounds intended to assess the overall quality after having completely analysed the text and reflected on every aspect of the translation or post-editing. They also had to guess whether the assessed document was a translation or a post-editing (post-editing/translation identification task). Evaluators were allowed to complete the assessments at their own pace during the 20 days allotted, as long as the correction of each document was done within the same day.

3.2.2. Evaluation Matrix and Questionnaire Design

The evaluation matrix used for this study is based on the Multidimensional Quality Metrics (MQM) established by Uszkoreit et al. (2013), as MQM is designed to assess both human and machine translations, and allows us to check only the categories that are considered relevant for each specific text type in as much or little detail as needed. MQM also permits the inclusion of other categories dealing with domain specific issues. Although MQM offers over one hundred categories and subcategories of issue types, only five categories and eleven subcategories were selected (see Table 2). The selection was based on previous research on the most common MT engines errors both in general texts (Avramidis et al., 2014) and in wildlife documentary films (Ortiz-Boix, forthcoming), and also in post-edited texts (Guerberof Arenas, 2009). Furthermore, as MQM does not include any audiovisual translation domain specific issues, a new category containing four subcategories was included: voice-over/off-screen dubbing specificities (see table 2). These specificities were: spotting[9], action synchrony, voice-over isochrony (Franco et al., 2010), and inclusion of phonetic transcriptions to facilitate the pronunciation of foreign names by voice talents. Although these elements would not be relevant when analysing machine translation output, they are considered in the current experiment, where the quality evaluation is done not on the machine translation output but on post-edited texts that have to fulfill certain requirements of the audiovisual transfer mode under review. Table 2 summarises the evaluation matrix used. The identification number that will be later used in the analysis for each category is included in the right column.

[9] Spotting, also called timing or cueing, is the process of defining in and sometimes out time codes of each voice-over or off-screen dubbing unit (Ortiz-Boix, forthcoming).

Issue Types Categories	Issue Types Subcategories	ID. Numbers
Adequacy	Wrong translation	1.1.
	Omission	1.2.
	Addition	1.3.
	Non-translated words	1.4.
Fluency	Register	2.1.
	Style	2.2.
	Inconsistencies	2.3.
	Spelling	2.4.
	Typography	2.5.
	Grammar	2.6.
	Others	2.7.
Variety		3
Voice-over specificities	Spotting	4.1.
	Action and kinetic synchronies	4.2.
	Phonetic transcriptions	4.3.
	Isochrony	4.4.
Design / Layout		5
Others		6

Table 2. *Evaluation matrix: error typology*

Evaluators were given an explanatory document with a definition and example of each issue type. Prior to the experiment, a pilot test and specific training were carried out to confirm the categories were appropriate and procedures were correctly understood.

As for the questionnaire, the participants had to report their level of agreement with eight statements on a 7-point Likert scale, where one equated to "completely disagree" through seven equating to "completely agree". The statements were the following:

(1) Generally speaking, the text was fluent;

(2) Overall, the text was grammatically correct;

(3) Broadly speaking, there were no spelling mistakes;

(4) Generally speaking, the vocabulary was appropriate;

(5) The vocabulary was mostly coherent throughout the text;

(6) In general, the text fulfills the standards of voice-over translation;

(7) Overall, the final result was satisfying;

(8) The text could be sent to a dubbing studio to be voiced-over.

3.2.3. Data and Methods

After all the evaluators performed their tasks, one hundred and forty-four corrected documents were collected along with the corresponding questionnaires. They were analysed using the statistical system R-3.1.2 (https://www.r-project.org/), developed by John Chambers and colleagues at Bell Laboratories. The following data were obtained and analysed:

(1) The grades for each document in the three scoring rounds. For round 1 (after reading the document for the first time) and round 2 (after correcting the documents, before answering the questionnaire), the following scale was used: "completely unsatisfactory", "deficient", "fail", "pass", "good", "very good" and "excellent". For round 3 (after correcting the documents and answering the questionnaire), a more precise scale similar to the ones lecturers apply in their courses' assessment was used: a numerical scale from 0 to 10, being 0 the lowest mark and any mark below 5 equal to a “fail”. The data for rounds 1 and 2 are discussed in Subsection 4.1 globally and in Subsection 4.3. separately for each excerpt. The data for round 3 can be found in Section 4.5.

(2) 144 questionnaires (6 x 24 documents) reporting on the participants' opinions after correcting each document. An analysis of the questionnaire replies is provided in Section 4.2. (globally) and 4.4 (separately for each excerpt).

(3) The results of the post-editing/translation identification task. This issue is discussed in Subsection 4.6.

(4) 144 documents with corrections according to the evaluation matrix based on the MQM: 6 corrected documents for each of the 24 documents. The number and type of errors corrected are discussed in Subsection 4.7.

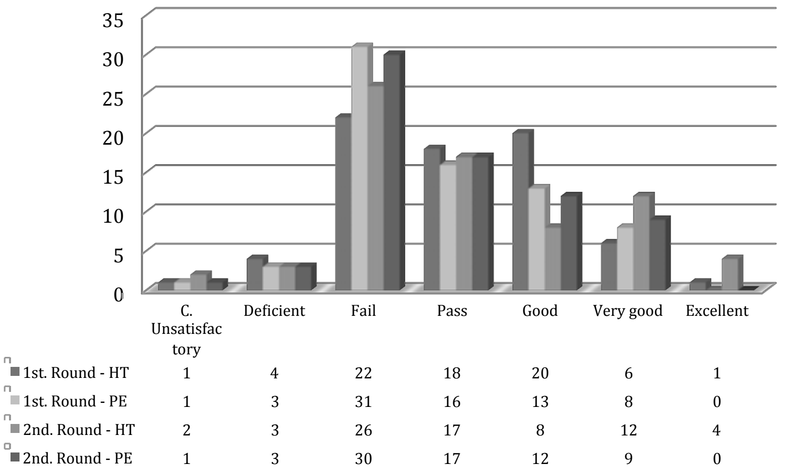
An ANOVA variance test was carried out to validate the hypothesis. Statistical significance is assumed for $p<0.05$, meaning that the difference between the results of the post-editing and translation QA should be higher than 0.05 to be considered significant.

3.3. Discussion of Results: Scoring Rounds and Questionnaire Replies

This Section discusses the global results taking into account scoring rounds 1 and 2 and the questionnaires replies (Subsection 1), and it then considers each excerpt separately (Subsection 2). Next, it presents the results of scoring round 3 (Subsection 3). It finally discusses the post-editing/translation identification task (Subsection 4).

3.3.1. Global Analysis

The results of the global analysis, which includes the data from both excerpts, indicate that the differences, in terms of quality, between translation and post-editing of wildlife documentary films are not high. Thus, these results seem to validate the hypothesis of the study, as evaluators consider translations and post-edited texts qualitatively comparable both in the case of the grades given in the scoring rounds (see Figure 1) and in the questionnaire-based evaluation (see Figure 16).



	C. Unsatisfactory	Deficient	Fail	Pass	Good	Very good	Excellent
■ 1st. Round - HT	1	4	22	18	20	6	1
■ 1st. Round - PE	1	3	31	16	13	8	0
■ 2nd. Round - HT	2	3	26	17	8	12	4
■ 2nd. Round - PE	1	3	30	17	12	9	0

Figure 1. *Global evaluation results: scoring rounds*

Furthermore, when focusing only on the evaluators' first scoring round, translations are better than post-edited texts, as 62.5% of translations (45 out of 72) have been evaluated from “pass” to “excellent” whilst only 51.38% of the post-edited texts have (37 out of 72). Furthermore, when focusing only in the best-rated outputs (from “good” to “excellent”), translations also get better scores (27 vs. 21). However, the median value for both translations and post-editing in round 1 is the same: “pass”.

As far as round 2 is concerned, the difference between translations and post-edited texts is reduced: 56.95% (41 out of 72) of the translations and 52.78% (38 out of 72) of the post- editings were evaluated from “pass” to “excellent”. When only considering those between “good” and “excellent”, 33.30% of the translations (24) and 29.17% of the post-edited texts (21) are found in this range. Even narrowing the scope to the best outputs, the number of translations that can be included in the ranges between “good” and “excellent” is higher than the number of post-edited texts included in them. However, when descriptive statistics are performed, it can be seen that the median grade for both tasks is “pass”.

Comparing the results of round 1 and 2, results are lower in round 2, which might lead to the conclusion than the more in depth the raters assess, the stricter they are and fewer differences between translation and post-editing are observed. The strictness of the second round might be due to the fact that evaluators had assessed the translated and post-edited documents in a more didactic way and according to a set of specifications. Hence, they could be aware of problems that had been not noticed during the first round, when they evaluated the translated and post-edited texts globally, adopting a more audience-centric perspective. Additionally, although the results of rounds 1 and 2 seem to indicate that translations are better than post-edited texts, the results are not statistically significant (1st round: F=0.000, p=1.000; 2nd round: F=1.000, p=1.000)[10], leading to the conclusion that post-edited texts and translations are significantly similar in terms of quality.

When adopting a different evaluation system, the results are slightly different. The questionnaire-based assessment indicates post-edited texts are better than translations: translations are given lower grades in 4 out of 8 specific evaluation issues – grammar, coherence, correction and adequacy of the text so that it can be sent to a dubbing studio- and the same grade in another item – VO specificities (see Figure 2).

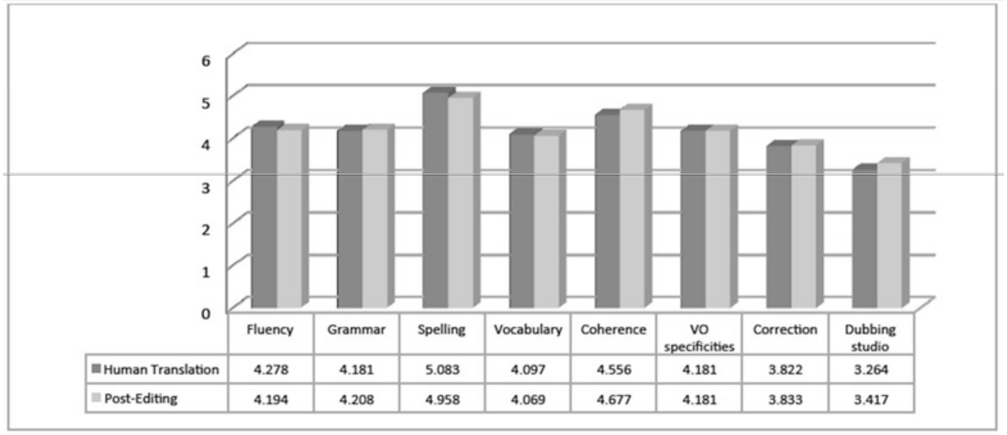


Figure 2. *Global evaluation results: questionnaire-based assessment (mean values)*

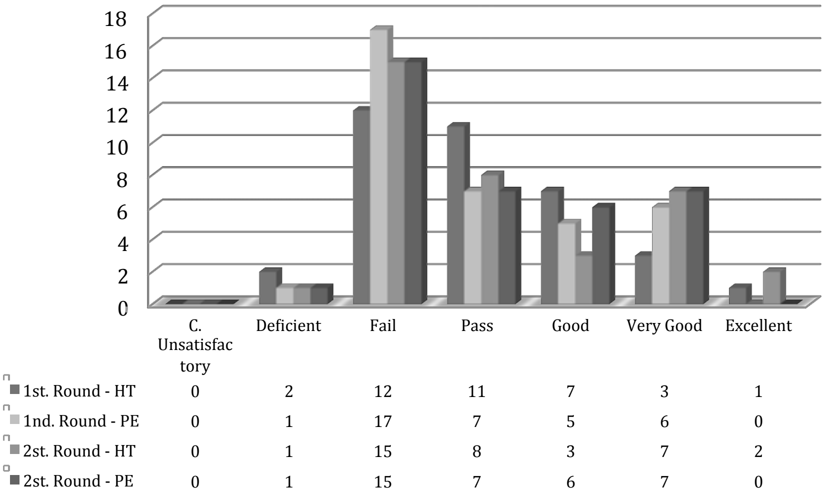
[10] "F=" stands for "F-value", which shows if a group of variables are significant together. "p=" stands for "p-value", which shows the provability of obtaining an equal or similar result to what has been observed in this particular experiment and, hence, its significance.

In this case, the evaluators scored the issues from 1 to 7, with 1 being the lowest grade. In those instances where translations got higher grades, mean grades for post-edited texts are no more than 0.1 points lower. The largest difference (0.247 points) can be found for issue type "coherence", where post-edited texts get a better grade. And the smallest difference is for "voice-over specificities", where there is no difference between translations and post-edited texts (4.181). However, in all categories the difference between post-edited texts and translations is again non-significant: fluency (F=0.155, p=0.695), grammar (F=0.004, p=0.948), spelling (F=0.691, p=0.407), vocabulary (F=0.019, p=0.892), coherence (F=0.410, p=0.523), VO specificities (F=0.000, p=1.000), correction (F=1.450, p=0.230), dubbing studio (F=0.581, p=0.447). When the questionnaire-based assessment is correlated with the global analysis, results indicate that issues related to terminological coherence, grammar and dubbing studio specificities have more impact on the grades than issues related to spelling, vocabulary or fluency, as the difference between translations and post-edited texts shortens after evaluation round 2.

Results show, therefore, that in general there is no significant difference between the quality of post-edited texts and translations in the analysed excerpts.

3.3.2. Specific Analysis

When the results for each excerpt are analysed separately, some differences appear, even though the results are not statistically significant here either. For excerpt 2, results indicate that the quality of post-editing and translation after evaluations rounds 1 and 2 is almost equal (see Figure 3).

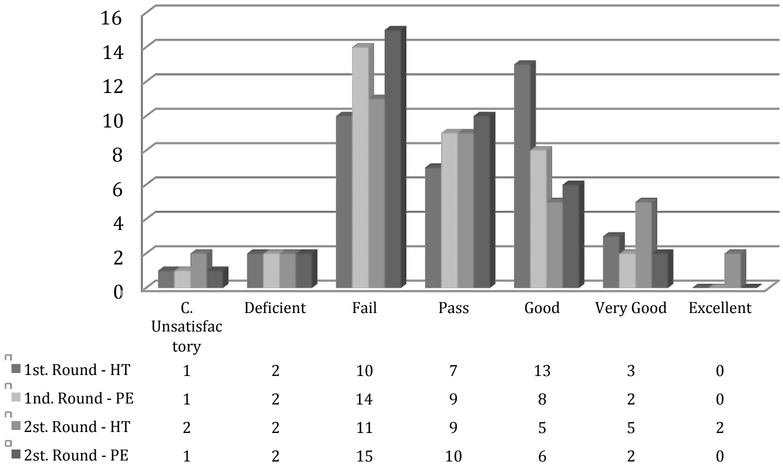


	C. Unsatisfactory	Deficient	Fail	Pass	Good	Very Good	Excellent
1st. Round - HT	0	2	12	11	7	3	1
1nd. Round - PE	0	1	17	7	5	6	0
2st. Round - HT	0	1	15	8	3	7	2
2st. Round - PE	0	1	15	7	6	7	0

Figure 3. *Specific evaluation results: scoring rounds (excerpt 2)*

In round 1, translations are slightly better than post-edited texts, as 61.11% of the translations (22) versus 50% of the post-edited texts (18) get a pass grade (from “pass” to “excellent”). However, when focusing on the outputs in the range between “good” and “excellent”, the number of post-edited texts and translations is the same (30.55%, 11 out of 36), and there are more post-edited texts rated as “very good” than translations (6 versus 1). However, the median for both translations and post-edited texts is again the same: “pass”. In round 2, if data are divided into two groups (fail and below/pass and above), the percentage is exactly the same for both post-edited texts and translations: 44.45% (16 out of 36) vs 55.55% (20 out of 36). However, when looking at the distribution in the higher range, one can observe that translations get higher marks than post-edited texts. Again, though, the median grade for both tasks is “pass”, showing no significant differences with round 1. It must be noticed that raters were stricter in round 2 and less differences were found between translations and post-edited texts. Results in both rounds were not statistically significant (round 1: F=0.584, p=0.447; round 2: F=0.004, p=0.748), confirming that post-edited texts and translations in our experiment are quite similar in terms of quality.

For excerpt 1, however, results present wider differences between post-edited texts and translations, as shown in Figure 4.



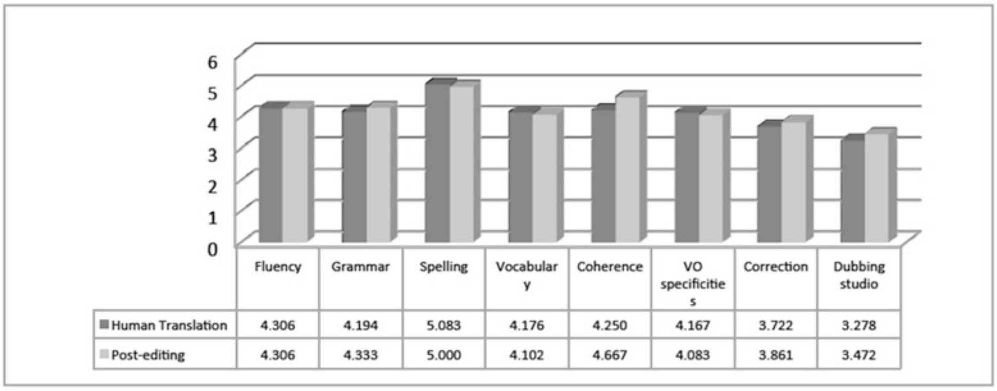
	C. Unsatisfactory	Deficient	Fail	Pass	Good	Very Good	Excellent
■ 1st. Round - HT	1	2	10	7	13	3	0
■ 1nd. Round - PE	1	2	14	9	8	2	0
■ 2st. Round - HT	2	2	11	9	5	5	2
■ 2st. Round - PE	1	2	15	10	6	2	0

Figure 4. *Specific evaluation results: scoring rounds (excerpt 1)*

In round 1, translations seem to be better than post-edited texts, as 63.89% (23 out of 36) get pass grades (from "pass" to "excellent") compared to 52.77% of post-edited texts (19). The difference is still more striking in the higher marks: 44.4% of translations (16 out of 36) get between "good" and "excellent", whilst only 27.78% (10 out of 36) post-edited texts are found in this range. However, statistics show that the median score for both conditions is the same: "pass". In round 2, the difference between translations and post-edited texts evens out, as 58.33% (21 out of 36) and 50% (18) of post-edited texts get pass grades. In the higher marks, the difference is 33.33% translations (12) versus 22.22% post-edited texts (8). The median grade for translation is "pass", whilst the median for post-editing falls between "pass" and "fail", again showing the tendency of evaluators to be stricter in second rounds. When inferential statistics are performed, results show again no significant differences in both rounds (round 1: F=0.584, p=0.447; second round: F=0.004, p=0.948), confirming the conclusion that post-edited texts and translations are significantly similar in terms of quality.

To summarize, although the results of the global analysis (including both excerpts) indicate that translations receive better marks than post-edited texts, the difference between them is not statistically significant. When the results are divided according to the excerpts, opposing trends are observed: in excerpt 2 post-edited texts receive better grades than translations, whilst in excerpt 1 translations receive higher marks. However, differences are not statistically significant in any of the excerpts under analysis. The results also indicate that the difference between translation and post-editing narrows after each round of evaluation.

As for the questionnaire responses, the biggest differences are found in the second excerpt (see figure 5), where post-editing received better grades than translation in four of the specific evaluation issues. However, these differences are not high, with coherence being the issue type where the widest difference is to be observed (4.25/4.667).

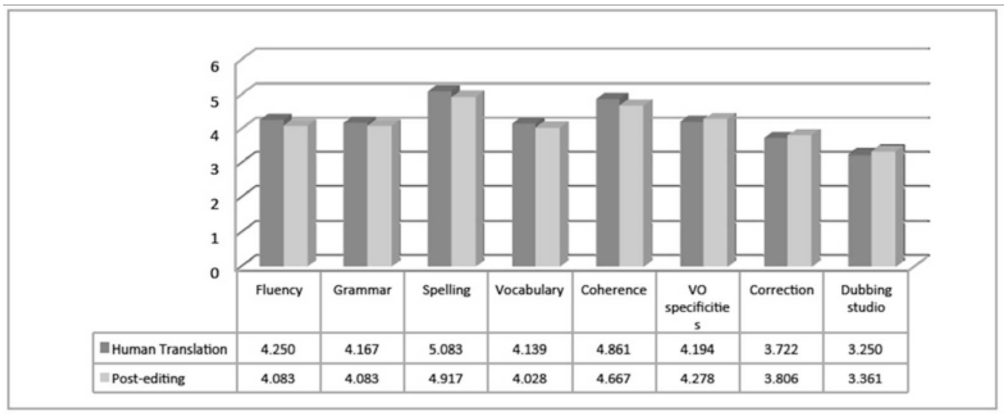


	Fluency	Grammar	Spelling	Vocabulary	Coherence	VO specificities	Correction	Dubbing studio
■ Human Translation	4.306	4.194	5.083	4.176	4.250	4.167	3.722	3.278
■ Post-editing	4.306	4.333	5.000	4.102	4.667	4.083	3.861	3.472

Figure 5. *Specific evaluation results: questionnaire-based assessment (excerpt 2)*

Although the grades for post-edited texts are generally higher than for translations, such differences are not statistically significant in any case: fluency (F=0.000, p=1.000), grammar (F=0.254, p=0.616), coherence (F=3.182, p=0.079), correction (F=2.248, p=0.138), and adequacy for the dubbing studio (F=0.506, p=0.479). When translations perform better, differences are again non-significant: spelling (F=0.103, p=0.749), vocabulary (F=0.042, p=0.839) and VO specificities (F=0.100, p=0.753).

The results for the first excerpt, however, again present much narrower differences (see Figure 6).



	Fluency	Grammar	Spelling	Vocabulary	Coherence	VO specificities	Correction	Dubbing studio
▨ Human Translation	4.250	4.167	5.083	4.139	4.861	4.194	3.722	3.250
▨ Post-editing	4.083	4.083	4.917	4.028	4.667	4.278	3.806	3.361

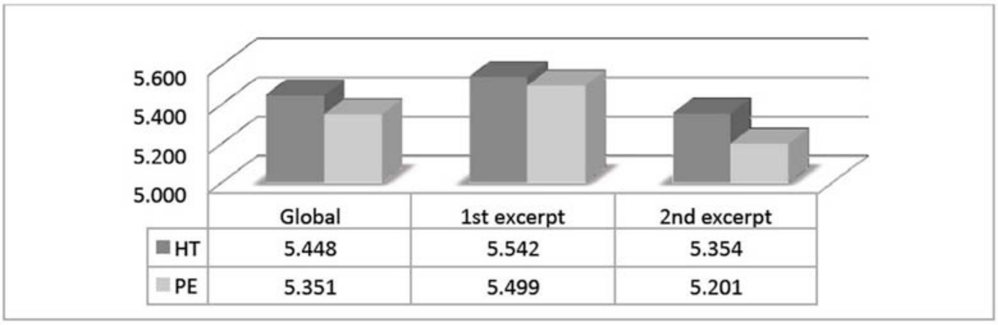
Figure 6. *Specific evaluation results: questionnaire-based assessment (excerpt 1)*

Translations are better than post-edited texts in five aspects, although the differences are not significant in any of the cases: fluency (F=0.266, p=0.608), grammar (F=0.267, p=0.607), spelling (F=0.735, p=0.394), vocabulary (F=0.130, p=0.719) and coherence (F=0.608, p=0.438). In other aspects, such as VO specificities (F=0.108, p=0.743), correction (F=0.079, p=0.780) and adequacy for the dubbing studio (F=0.140, p=0.709), post-edited texts are evaluated as better. The difference between the mean grades is not significant in any aspect, which leads to the belief that translation and post-editing are comparable.

To sum up, in all cases, translation and post-editing are significantly similar. Although there are differences between translations and post-edited texts in both excerpts, such differences are minimal and, therefore, the results prove the null-hypothesis of the article is correct.

3.3.3. Third Round of Evaluation

After answering the questionnaire, evaluators graded each text for the last time (round 3). In this case, they gave translation and post-edited texts a grade from 0 to 10, being 0 the lowest grade and 10 the highest, as this is the scale lecturers use to evaluate at university. Figure 7 presents the mean results of this evaluation.



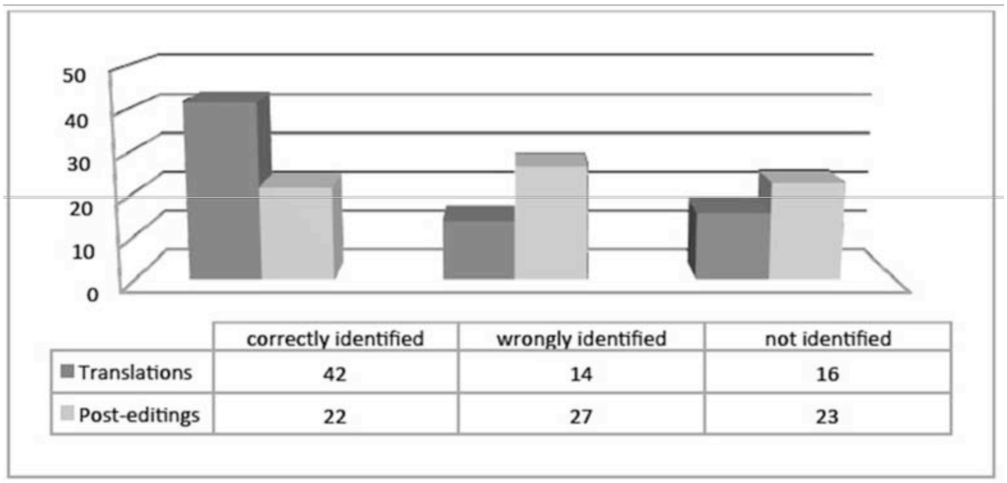
	Global	1st excerpt	2nd excerpt
▨ HT	5.448	5.542	5.354
▨ PE	5.351	5.499	5.201

Figure 7. *Evaluation round 3*

As far as the global final grade (including both excerpts) is concerned, mean grades are 5.3505 for post-editing and 5.448 for translation, a difference which is not statistically significant (F=0.000, p=1.000). When each excerpt is analysed independently, results are also better for translations, although the difference is higher for the second excerpt (0.153) than for the first (0.043). Differences are again not statistically significant (F=0.000, p=1.000; F=0.001, p=0.975), which is in line with results in previous scoring rounds. Similarly, the slight differences found between translations and post-editing narrow in each of the evaluation rounds, with the difference between round 1 and 2 the highest.

3.3.4. Identification of Post-edited Texts and Translations

Evaluators were asked to identify each corrected document as a translation or as a post- editing. The results show that it is easier to assert which ones are translations (see Figure 8), as 42 out of 72 translations (58.33%) were correctly identified and only 14 translations (19.44%) were wrongly identified as post-edited texts. The remaining 16 translations (22.22%) were not identified by the evaluators, who indicated on the form they did not know whether they were a translation or a post-editing.



	correctly identified	wrongly identified	not identified
■ Translations	42	14	16
□ Post-editings	22	27	23

Figure 8. *Identification of translations*

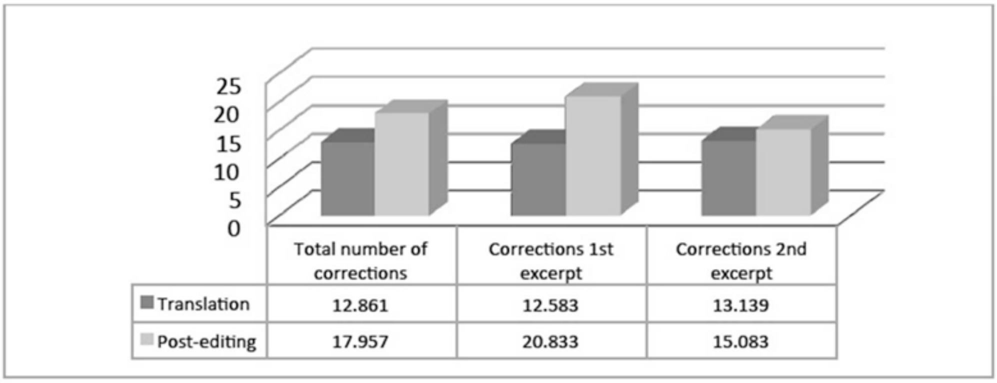
As for post-edited texts, they were more difficult to identify: while only 22 out of 72 post- edited texts (30.55%) were correctly identified as such, 27 documents were misidentified as translations (37.5%), and in 23 cases (31.94%), evaluators could not be sure of the text-types.

Summing up, although most translations are correctly identified, it seems that post-edited texts are difficult to identify as such, as the great majority of them are either misidentified or not recognized. These results may imply that the quality of post-edited texts can be considered comparable to the quality of translations. However, it remains to be seen to what extent the lack of experience of the evaluators with post-editing may have influenced the results and whether an explicitly mentioned revision task made by the same translators after the translation task would increase the quality of the translations. It should be noticed that translators were instructed to provide a translation that would be fit for recording; hence an implicit revision task was included but not verbalized in the instructions.

4. Discussion of Results: Correction based on the Evaluation Matrix

This section analyses in detail the assessments made by evaluators, focusing on the number and type of mistakes found in the translations and post-edited texts, both globally and separately for each excerpt.

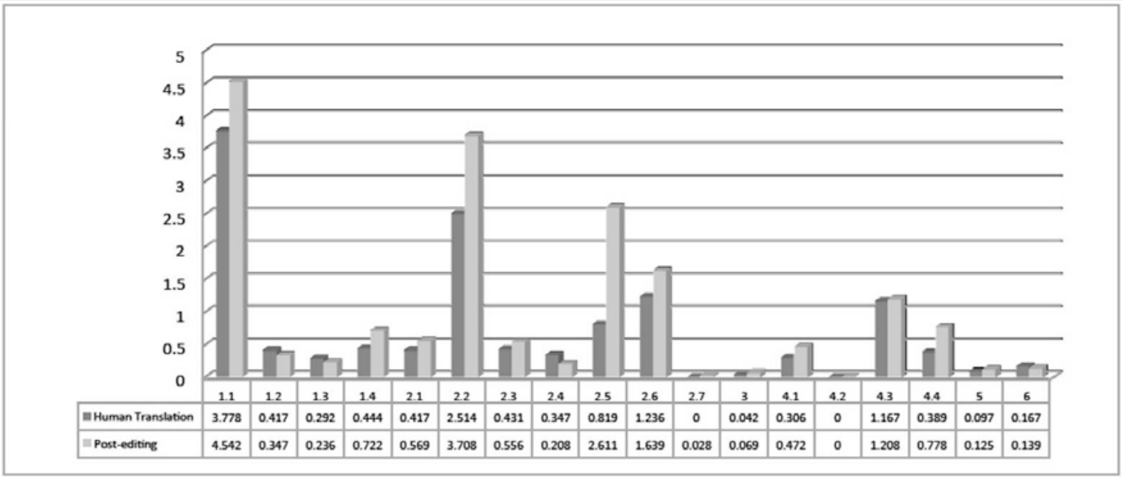
Results show that translations needed in general a lower number of corrections than post- editings. The mean difference is 5 errors (see Figure 9), which indicates that the quality of translations and post-edited texts of wildlife documentary scripts could be considered similar. However, when the corrections are analysed separately for each excerpt, it can be observed that there is a considerable difference between the translation quality and the post-editing quality of the first excerpt (see Figure 7), as post-editing almost doubled the mean number of errors of the translation. As far as the second excerpt is concerned, the mean number of corrections in translations and post-edited texts is narrowed (see Figure 9).



	Total number of corrections	Corrections 1st excerpt	Corrections 2nd excerpt
■ Translation	12.861	12.583	13.139
■ Post-editing	17.957	20.833	15.083

Figure 9. *Number of corrections*

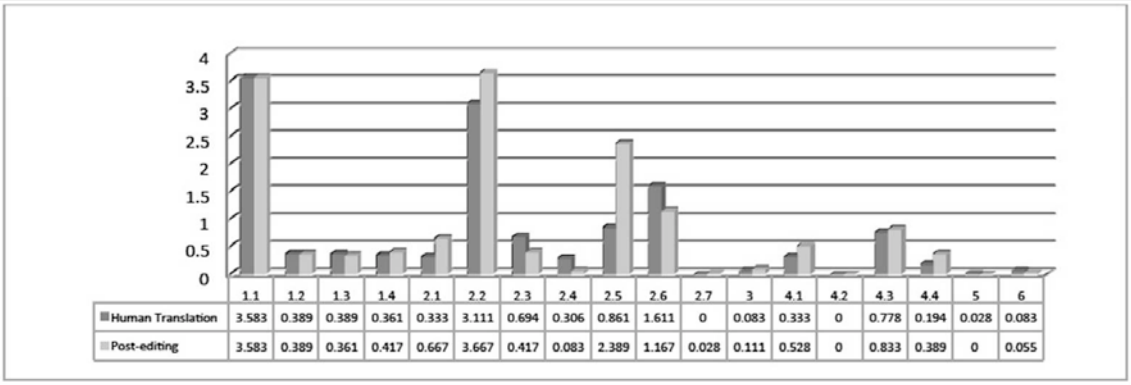
According to the global analysis, the mean number of errors within every issue type is similar (see Figure 10), with 1.195 corrections (2.2. style) the widest difference.



	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3	4.1	4.2	4.3	4.4	5	6
■ Human Translation	3.778	0.417	0.292	0.444	0.417	2.514	0.431	0.347	0.819	1.236	0	0.042	0.306	0	1.167	0.389	0.097	0.167
■ Post-editing	4.542	0.347	0.236	0.722	0.569	3.708	0.556	0.208	2.611	1.639	0.028	0.069	0.472	0	1.208	0.778	0.125	0.139

Figure 10. *Global evaluation: error typology (both excerpts)*

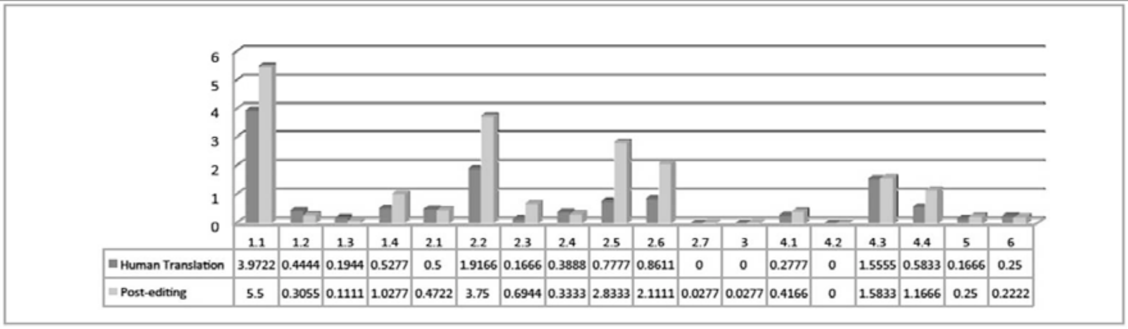
Figure 10 indicates that no corrections were performed on issue type 4.2. (synchrony). Furthermore, it shows that post-edited texts contain more errors in all issue types except in 1.2. (omission), 1.3. (addition), and 2.4. (spelling). Nevertheless, as has been observed in previous analyses, there is a change of tendency in the second excerpt. In this case, translations contain as many or more corrections in 8 issue types: 1.1. (wrong translation), 1.2. (omission), 1.3. (addition), 2.3. (inconsistencies), 2.4. (spelling), 2.6. (grammar), 5 (design/layout) and 6 (others).



	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3	4.1	4.2	4.3	4.4	5	6
■ Human Translation	3.583	0.389	0.389	0.361	0.333	3.111	0.694	0.306	0.861	1.611	0	0.083	0.333	0	0.778	0.194	0.028	0.083
■ Post-editing	3.583	0.389	0.361	0.417	0.667	3.667	0.417	0.083	2.389	1.167	0.028	0.111	0.528	0	0.833	0.389	0	0.055

Figure 11. *Specific evaluation: error typology (excerpt 2)*

The greatest difference between the number of corrections in translations and post-edited documents is found in issue type 2.5. (typography), being 1.5277 points. Compared to the global analysis, excerpt two has two issue types that contain as many corrections (3.583 and 0.389) for translations as for post-edited texts: 1.1. (wrong translation) and 1.2 (omission). In the case of the first excerpt, results are presented in figure 12.



	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3	4.1	4.2	4.3	4.4	5	6
■ Human Translation	3.9722	0.4444	0.1944	0.5277	0.5	1.9166	0.1666	0.3888	0.7777	0.8611	0	0	0.2777	0	1.5555	0.5833	0.1666	0.25
■ Post-editing	5.5	0.3055	0.1111	1.0277	0.4722	3.75	0.6944	0.3333	2.8333	2.1111	0.0277	0.0277	0.4166	0	1.5833	1.1666	0.25	0.2222

Figure 12. *Specific evaluation: error typology (excerpt 1)*

Although results are similar to those of the global analysis, the difference between the corrections performed in issue types 1.1. (wrong translation), 2.2. (style), 2.5 (typography) and 2.6. (grammar) is much wider (2.383, 2.233, 2.133 and 1.250 points respectively) in favor of translation. Moreover, five issue types contain more corrections for translation than post-editing: 1.2 (omission), 1.3 (addition), 2.1 (register), 2.4 (spelling) and 6 (others). No issue type has as many corrections for translating as for post-editing.

On the one hand, results indicate that the errors contained in the translations are more varied, as the subcategories with more errors are from both the categories of accuracy and fluency, as well as domain specific issues: they contain more incorrectly translated words (issue type 1.1) than post-edited texts and more problems regarding register (type 2.1), typography (type 2.5), spotting (type 4.1) and phonetic transcriptions (type 4.3). On the other, results show that post-edited texts usually present more errors in style (type 2.2), grammar (type 2.6) and typography (type 2.5). Moreover, post-edited texts present fewer domain-specific errors in the domain of wildlife documentary films. Thus, it can be observed that, as in other studies that assess post-edited texts versus translations (see Section 2), the quality of post-edited texts is lower only with regards to fluency, which indicates that MT might help with accuracy issues and might allow translators to focus on domain specific issues. Furthermore, it also leads us to believe that better results for the post-edited texts are likely to be obtained through using an MT engine built with in-domain data instead of a free online MT engine like Google Translate, as other errors might be avoided and more terminology could be included.

5. Summary and Conclusions

This study is built upon the hypothesis that the quality of post-edited and translated wildlife documentary films is significantly similar and proves its null-hypothesis. The results are presented both globally and separately for each excerpt according to the various types of evaluation data obtained: three scoring rounds, questionnaire-based evaluation, post-editing/translation task identification, and evaluation matrix-based assessment.

The results of all evaluation systems, both globally and separately for each excerpt, correlate and prove that there are no significant differences between post-editing and translation in terms of quality, hence validating the null-hypothesis of the study. Although non-significant, it must be stressed that the differences between translations and post-edited texts vary depending on the excerpt: while translation achieves better results in the first fragment, post-editing has higher marks in the second. Such differences between excerpts might be due to slight differences in their complexity.

When analysing the results of the questionnaire responses, it can be observed that post-edited texts are generally assessed more positively for terminology coherence and domain specific issues, whilst translations are graded better for fluency and general vocabulary. This might indicate that, as post-editors have accuracy issue types solved, they can focus on other issue types, such as domain specific problems and terminological coherence of the text.

As for the correction of the documents based on the evaluation matrix, the results show that the most common errors in texts translated by humans differ from those in post-edited texts: while post- edited texts have many errors regarding style and grammar, translations have more errors regarding mistranslated words. Thus, correction results correlate with the results of the questionnaire, as the most common errors produced in post-edited texts fall into the fluency category, and the most common errors found in translations fall into the accuracy and domain specific categories. As mentioned before, such results might be due to the fact that MT helps with accuracy issues and might allow translators to focus on domain specific issues.

Finally, the results of the translation/post-editing identification task show that evaluators are only able to identify a third of the post-edited texts. If post-edited texts were expected to be significantly different in terms of quality, one could expect a higher number of post-edited texts to be identified as such, but this is not the case. This compels us to claim once again that the quality of translations and post-edited texts in our experiment can be considered similar. However, it is not clear whether the experience of evaluators in terms of post-editing might have influenced such results.

All in all, the results seem to indicate that there are no significant differences between the quality of post-edited texts and translations of the wildlife documentaries used in our experiment. Results also show that the quality of both the translations and the post-edited texts was considered to be low, as the highest mean grade is just a few points above 5, the minimum pass grade in the Spanish system. It remains to be seen whether greater differences would be found in higher quality outputs.

Further research with other language pairs and a higher number of judges should be carried out to confirm the results, because the study is limited to one language pair (English into Spanish) and six human judges. Furthermore, it would be interesting to do similar testing with translations and post-editing produced by experienced translators and post-editors, since in our experiment translators and post-editors were volunteer MA students who had almost no previous professional experience and, consequently, the overall quality was affected by this. Another step would be to research audience reception; in other words, to test how TV audiences would receive a translated wildlife documentary versus a post-edited wildlife documentary. Many possibilities emerge, but this article has hopefully been another step towards future studies on the implementation of translation technologies in the field of audiovisual translation and media accessibility, an area that is still under-researched especially where oral modalities such as voice-over or dubbing are concerned.

References

Armstrong, S., Caffrey, C., Flanagan, M., Kenny, D., & Way, A. (2006). Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation. In *MuTra-Multidimensional Translation Conference Proceedings. Audiovisual Translation Scenarios* (no page numbers). Copenhagen: MuTra. Retrieved from http://www.mt-archive.info/Aslib-2006-Armstrong.pdf [Consulted: 29 June 2016].

Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S., & Sadler, L. (1994). *Machine Translation: An Introductory Guide.* (p. 240). Oxford: NCC Blackwell.

Avramidis, E., Burchardt, A., Federmann, C., Popovic, M., Tscherwinka, C., & Torres, D. V. (2012). Involving Language Professionals in the Evaluation of Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 1127-1130). Istanbul: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/294_Paper.pdf [Consulted: 13 July 2016]

Aziz, W., de Sousa, S., & Specia, L. (2012). PET: a Tool for Post-Editing and Assessing Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 3982-3987). Istanbul: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf [Consulted: 29th June 2016]

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. (2011). The process of postediting: a pilot study. In *Proceedings of the 8th international NLPSC workshop* (p. 131-142). Frederiksberg: Samfundslitteratur. Copenhagen Studies in Language.

Carroll, J.B. (1966). An Experiment in Evaluating the Quality of Translation. *Mechanical Translation*, *9*(3-4), 55-66.

Dabbadie, M., Hartley, A., King, M., Miller, K.J., El Hadi, W.M., Popescu-Belis, A., Reeder, F., & Vanni, M. (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. In *Proceedings of the Workshop at the LREC 2002 Conference: Machine Translation Evaluation: Human Evaluators Meet Automated Metrics* (8, p. 8-16). Las Palmas: ELRA. Retrieved from https://www.semanticscholar.org/paper/Workshop-at-the-Lrec-2002-Conference-Machine-Palmas/7a4d645f83068c16c919289791c7fe7759677765/pdf [Consulted: 13 July 2016]

Etchegoyhen, T., Bywood, L., Georgakopoulou, P., Fishel, M., Jiang, J., Loenhout, G., Pozo, A., Turner, A., Volk, M., & Maucec, M. (2014). Machine Translation for Subtitling: A Large-scale Evaluation. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)* (p. 46-53). Reykjavik: ELRA. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/463_Paper.pdf [Consulted: 13 July 2016]

Federmann, C., Melero, M., Pecina, P., & van Genabith, J. (2012). Towards Optimal Choice Selection for Improved Hybrid Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, *97*(1), 5–22. Retrieved from http://www.degruyter.com/view/j/pralin.2012.97.issue--1/v10108-012-0001-1/v10108-012-0001-1.xml [Consulted: 13 July 2016]

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

Franco, E., Matamala, A., & Orero, P. (2010). *Voice-over Translation: an Overview* (p. 249). Bern: Peter Lang.

Gambier, Y. (2008). Recent developments and challenges. D. Chiaro, C. Heiss, C. Bucaria (eds). *Between text and image: Updating research in screen translation* (78, p. 11-34). Amsterdam/Philadelphia: John Benjamins.

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

Giménez Linares, J.Á. (2008). *Empirical machine translation and its evaluation*. (PhD Dissertation). Universitat Politècnica de Catalunya.

Guerberof Arenas, A. (2009). Productivity and quality in MT post-editing. In *Proceedings of the MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT* (no pages). Ottawa: AMTA. Retriebed from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.575.5398&rep=rep1&type=pdf [Consulted: 13 July 2016]

—. (2012). *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. (PhD Dissertation). Universitat Rovira i Virgili.

House, J. (2006). Text and context in translation. *Journal of Pragmatics*, *38*(3), 338-358.

Hutchins, J., & Somers, H. (1992). *An Introduction to Machine Translation* (p. 362). London: Academic Press.

King, M., Popescu-Belis, A., & Hovy, E. (2003). FEMTI: Creating and Using a Framework for MT Evaluation. In *Proceedings of Machine Translation Summit IX* (p. 224-231). New Orleans: AMTA.

Koller, W. (1987). Zum Gegenstand der Übersetzungstheorien. In R. Arntz, & G. Thome (eds). *Übersetzungswissenschaft. Ergebnisse und perspektiven* (p. 19-30). Tübingen: Narr.

Krings, H.P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*, 5. Kent: Kent State University Press.

Kupsch-Losereit, S. (1988). Die Übersetzung als soziale Praxis. Ihre Abhängigkeit vom Sinn- und Bedeutungshorizont des Rezipienten. *Fremdsprachen Lehren und Lernen*, *17*, 203-216.

Lavie, A., & Denkowski, M.J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105-115.

Lommel, A. (ed). (2014) *Multidimensional Quality Metrics (MQM) Specifications*. Retriebed from <http://www.qt21.eu/mqm-definition/mqm-spec-2014-02-14.html> [Consulted: 29 June 2016] http://www.qt21.eu/mqm-definition/mqm-spec-2014-02-14.html

Matamala, A., Fernández-Torné, A., & Ortiz-Boix, C. (2012). Technology and AD: The TECNACC Project. In *Languages and the Media 2012 Conference*. Berlin: ICWE. Retrieved from http://ddd.uab.cat/pub/presentacions/2012/117159/fernandez_matamala_ortiz_berlin2012.pdf [Consulted: 16 May 2016]

Nida, E.A. (1964). *Toward a Science of Translation: with special reference to principles and procedures involved in Bible translating* (p. 331). Leiden: Brill.

Nida, E.A. & Taber, C.R. (1969). *The Theory and practice of Translation* (p. 229). Leiden: Brill.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, *19*(1), 37-58.

Ortiz-Boix, C., & Matamala, A. (forthcoming). Post-Editing Wildlife Documentary Films: a new possible scenario? *Journal of Specialized Translation (JosTrans), 26*, 187-210.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (p. 311-318). Stroudsburg: ACL. Retreived from http://delivery.acm.org/10.1145/1080000/1073135/p311-papineni.pdf?ip=176.85.32.252&id=1073135&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=643378208&CFTOKEN=84140615&__acm__=1468451020_756ba83ab9f75e6b1d06a593338750ed [Consulted: 13 July 2016]

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

Popović, M., & Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, *37*(4), 657-688.

Reiss, K. (1978). Übersetzungstheorie und Praxis der Übersetzungskritik. In F. Königs (ed). *Übersetzungswissenschaft und Fremdsprachenunterricht* (p. 71-93). München: Goethe-Institut.

Reiss, K., and Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie* (p. 254). Tübingen: Niewmeyer.

Roturier, J. (2006). *An Investigation into the impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical.* (PhD Disseratation). Dublin City University.

Savoy, T. (1968). *The Art of Translation* (p. 322). Boston: The Writer.

Snover, M., Dorr, B.J., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas* (p. 223-231). Cambridge: AMTA.

Snover, M., Madnani, N., Dorr, B.J., & Schwartz, R. (2009). Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (p. 259-268). Athens: ACL.

Sousa, S., Aziz, W., & Specia, L. (2011). Assessing the Post-editing Effort for Automatic and Semi-automatic Translations of DVD subtitles. In *Proceedings of Recent Advances in Natural Language Processing Conference* (p. 97-103). Hissar: ACL. Retrieved from http://aclweb.org/anthology/R11-1014 [Consulted: 29 June 2016].

Sutter, N., & Depraetere, I. (2012). Post-edited translation quality, edit distance and fluency scores: report on a case study. Paper presented at the *Journée d'études.*

Toury, G. (1985). *In Search of a Theory of Translation*. Tel Aviv: Tel Aviv University.

Uszkoreit, H., & Lommel, A. (2013). *Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment*. Retrieved from http://www.qt21.eu/launchpad /sites/default/files/MQM.pdf [Consulted: 29 June 2016].

Chapter 5. Fourth Article

5. Fourth Article

Ortiz-Boix, C. and Matamala, A. (2015). Assessing the Quality of Post-Edited Audiovisual Products: wildlife documentary films in voice-over and off-screen dubbing. *Proceedings of the 4th Workshop on Post-editing Technology and Practice (WPTP).* 16-30.

ABSTRACT: This article presents the results of a study on quality assessment designed to evaluate the quality of post-edited – in comparison to translated – wildlife documentary films which are delivered using voice-over and off-screen dubbing. The study proposes a quality assessment at three levels: raters assessment, in-studio assessment and users' assessment. The main contribution of this QA proposal is the inclusion of the end-users – via a user reception study – to the process of assessing the quality of a post-edited or translated audiovisual text. Results show that there is no significant difference between the quality of post-edited and translated wildlife documentary films.

Credits

This article is part of the ALST project (reference FFI-2012-31024) led by Anna Matamala and funded by the Spanish "Ministerio de Economía y Competitividad" and it is also partially funded by research group TransMedia Catalonia (2014SGR27) and the grant FI-DGR2013 awarded by AGAUR. This article is part of the research carried out by Carla Ortiz-Boix under the supervisions of Dr. Anna Matamala within the PhD in Translation and Intercultural Studies from the Department of Translation, Interpretation and East Asian Studies at Universitat Autònoma de Barcelona.

1. Introduction

Quality and quality assessment (QA) have been a main issue in Translation Studies since the academic discipline began in the late 1950s. Many studies have been carried out in that regard (e.g. Nida, 1964; Reiss et al., 1984; Gambier, 1998; Hansen, 2008; Melby, 2014) and have approached both quality and QA distinctively depending on the translation theory each QA model is based on (House, 2006). Studies on machine translation (MT) and post-editing (PE) have also addressed quality and QA creating models and measures to evaluate the quality of the types of texts (technical and general) more frequently applied to MT and PE. Although recent studies (Melero et al., 2006; Bywood et al., 2012; Etchegoyhen et al., 2014; Fernández-Torné et al., 2013; Ortiz-Boix et al., forthcoming) have proved the possibility to include MT and MT plus PE into the workflow of some audiovisual translation (AVT) modalities, mostly subtitling and especially in terms of productivity, more research into the topic of quality and QA of both MT and PE in AVT is still needed.

This article presents a proposal of QA to assess the quality of post-edited wildlife documentary films delivered through voice-over (VO) and off-screen dubbing (OD) transfer modes, as their specificities vary from those that can be found more frequently in the studies on QA of machine translated and post-edited texts. The proposal introduces a new level to the QA that has been usually left aside: quality assessment by the end-users of the final product, by means of a user reception study. Furthermore, it includes a brief quality assessment by a dubbing studio, which recorded the translated and post-edited versions that were used afterwards in the user reception test.

After the introduction the article is divided in five more sections. The following section (Section 3) briefly describes previous work on voice-over and off-screen dubbing, post-editing QA, and QA in AVT. Section 4 describes the QA proposal, and section 5 specifies the methodology used to implement the proposal. In section 6, results are presented. Finally, conclusions and further research are discussed in section 7.

2. Previous Work

This section defines voice-over and off-screen dubbing, highlighting the specificities of these AVT modalities (3.1). It then describes briefly the previous work on post-editing QA that has inspired the study (3.2), as well as previous work on QA in AVT (3.3).

2.1. Voice-Over and Off-Screen Dubbing

Voice-over (VO) is the AVT transfer mode that revoices an audiovisual text in another target language on top of the source language voice, so that both voices are heard simultaneously (Franco et al., 2010). In countries such as Spain, VO is the transfer mode frequently used in factual programs, e.g. documentary films, as it is said to help reproduce the feeling of reality, truth and authenticity given by the original audiovisual product due to visual and verbal evidences (Franco et al., 2010). In Eastern Europe, however, VO can also be found in fictional TV programs.

Off-screen dubbing (OD), also termed commentary and narration (Pönniö, 1995), is the transfer mode that revoices off-screen narrations substituting the original voice with a version in the target language (Franco et al., 2010). In other words, only the target language version is heard, not the original one. OD is used in factual programs and usually combined with VO (OD for off-screen narrators, VO for interviews).

Some of the main features of these transfer modes are the following:

(1) VO and OD have synchronization constraints. In VO three types of synchrony are observed: kinetic synchrony – the translating voice matches the body movements seen on screen–, action synchrony – the translating voice matches the actions seen on screen–, and voice-over isochrony – the translated message fits between the beginning and the end of the original speech, leaving some time after the original voice starts and before it ends–. OD is only endowed with kinetic and action synchronies, as the original voices are not heard in this transfer mode (Orero, 2006; Franco et al., 2010).

(2) Different language registers can coexist in the original product in which VO and OD are used: whilst VO is generally used for semi-spontaneous or spontaneous interviews, OD is usually applied to narrators with a planned discourse (Matamala, 2009; Franco et al., 2010). If the original product contains oral features such as fluffs, hesitations and grammatical mistakes, the target language version does not generally reproduce them (Matamala, 2009). In other words, the translation is generally an edited version of the original.

When translating documentary films, translators have to deal with the terminology of the subject matter of the program (Matamala, 2009). It is also often the case that the source text contains linguistic errors, inaccuracies and inconsistencies (Franco et al., 2010) and that a quality written script is not available (Ortiz-Boix, forthcoming). However, translators are expected to deliver a quality written script in the target language so that the recording by voice talents in a dubbing studio can begin.

2.2. Post-Editing Quality Assessment

Although research on QA of post-edited text has increased, it is still rather limited. Fiederer and O'Brien (2009), Plitt and Masselot (2010), Carl et al. (2011), García (2011), Guerberof Arenas (2009, 2012), Melby et al. (2014) and Mariana (2014) have dealt with quality in post-editing, to a greater or lesser extent. Up until now, QA has been based mostly on either holistic approaches –which assess the quality taking into account the whole text–, analytic approaches –which assess the quality of the text by analyzing the text in detail according to different sets of specifications – or, more recently, a combination of both.

Holistic approaches: Plitt and Masselot (2010) used the Autodesk translation QA team to assess randomly selected samples of translated and post-edited text in two labels "average" or "good", depending on whether they considered the text were fit for publishing the possibility to publish the texts they read according to their own subjective opinion. Carl et al. (2011) used a holistic approach as, even though raters assessed single sentences and not whole texts, assessors had to rank the same sentences either translated or post-edited according to their quality (ties were allowed). Fiederer and O'Brien (2009) also assessed the quality of sentences – three translated and three post-edited versions of 30 sentences – according to clarity, accuracy and style by ranking them from 1 to 4, being 1 the lowest mark. Raters were also asked to indicate their favorite option out of the six proposals for each source sentence.

Analytic approaches: In García (2011), a rater assessed the quality of a 500-word text by using the Australian National Accreditation Authority for Translators and Interpreter's (NAATI) guidelines. In Guerberof Arenas (2009, 2012) three raters blindly assessed translated segments, post-edited segments and segments previously extracted from a translation memory by using the LISA QA model, which has eight categories: mistranslation, accuracy, terminology, language, style, country, consistency and format.

Mixed approaches: Melby et al. (2014), Mariana (2014) and Lommel et al. (2014) present the Multidimensional Quality Metrics (MQM). The model provides a framework for defining metrics and scores used to assess the quality of human translated, post-edited or machine translated texts by setting error categories, otherwise called issue types, used to assess different aspects of the quality and identify problems. MQM are based on the translations specifications (Melby, 2014) that define the expectations of a particular type of translation. MQM is organized in a hierarchic tree that can have the issue types necessary for the type of text and the set of specifications that need to be assessed.

2.3. Quality Assessment in Audiovisual Translation

QA research in AVT is still rather limited and has mainly focused on subtitling (e.g. Gottlieb, 2001; Díaz Cintas, 2001) and dubbing (e.g. Chiaro, 2008). There are some norms (UNE 153010 or UNE 153020) that might be followed to fulfill a supposed minimum quality level for some AVT modes, VO and DO have none. However, QA of these two modes might follow more general standards, e.g. ISO Standards 9000 and 9001 on quality and quality management, and ISO Standard

EN 15038 and ISO 17100:2015 for quality in translation services. Such standards focus on the end-user and include benchmarking systems of quality management in relation to customer requirements and satisfaction. Hence, as the research on quality in AVT is limited and taking into account the importance of the end-user observed in ISO quality standards and in AVT, a functionalist approach to QA for post-edited wildlife documentary films is proposed. In the presented study, the end-user satisfaction is tested via a user reception study, as we believe it can help determine whether the purpose of the translation has been achieved: if a translation or post-editing meets its purpose, users react positively –e.g. they will enjoy and comprehend the product– and thus, the customer –or user– requirements, the translated or post-edited version can be considered a quality product.

In AVT there are more and more user reception studies. Gambier (2009), for example, presented the levels necessary to assess user reception based on Kovačič (1995) and Chesterman (2007). Accordingly, reception is assessed in three levels:

(1) Response or lisibility –which stands for the perceptual decoding by the user–. Sociological and audiovisual specific variables (e.g. age, level of education, aptitudes, hearing/sight difficulties, film genre or interplay images and dialogue) must be taken into account in this level.

(2) Reaction or readability –a psycho-cognitive issue that determines whether the previous knowledge of the users affects the reception and the inference process (understandability) of the viewers when watching the product, as the greater the processing effort is, the lower the relevance of the translation is. It also determines how the users react to the AVT product.

(3) Repercussion –both an attitudinal issue that tests the viewer's preferences and habits regarding the assessed AVT transfer mode and

the sociocultural dimension of the non-TV context which influences how the user process of reception.

3. Quality Assessment for Post-Edited Wildlife Documentary films

Taking into account the QA models presented in the previous section and the importance of the dubbing studio in oral AVT transfer modes, we build a proposal to assess the quality of post-edited wildlife documentary films in comparison to translated ones. This QA is done in three levels that include: QA by professional audiovisual translators to rate the translations and post-editings, QA by the dubbing studio and QA by users to rate their reception.

3.1. Quality Assessment by Raters[11]

QA by raters is the first level of the proposal, where raters perform a quality assessment both from a holistic and an analytic point of view.

Holistic QA: Raters are to give scores to each translated or post-edited document as a whole in three rounds. During the first round, raters give a score to the translation or post-editing as a whole according to their first impression without knowing which documents are translated or post-edited. During the second round, which is done after the analytic QA, raters give a score to the translation or post-editing as a whole according to their impression after analyzing the document in detail. The third scoring round is done at the end the QA, after answering a questionnaire on more concrete aspects and after deciding whether the document is a translation or a post-editing. The holistic QA also includes a questionnaire on 8 important parameters the translation or post-editing should met in order for it to be of good quality: general fluency, grammar, spelling, vocabulary appropriateness, terminology coherence, voice-over specification standards, satisfaction for the final result and purpose meeting; as well as a categorization of the document as "translation", "post-editing" or "any of

[11] See Ortiz-Boix and Matamala (forthcoming) for further details on this level of the QA proposal.

the above". The questionnaire is considered holistic, as raters are to evaluate the parameters in the whole document in general.

Analytic QA: Raters are to correct the text according to 18 parameters based on MQM and give weights to the parameters, if wanted. Such parameters can vary depending on the AVT transfer mode used (and example of possible parameters can be found in Table 1 - see section 5.4.).

3.2. Quality Assessment by the Dubbing Studio

Dubbing Studio does an analytic QA by making changes into the script so that it fully meets their requirements in order to be a quality product taking into account the particularities of VO and OD. Furthermore, the researchers take observational notes on the changes and problems – e.g. cacophonies or length of the sentences – detected in the recording of the voices. In the case of the present study, dubbing studio took into account voice-over isochrony as well as action and kinetic synchronies when making any changes on the scripts and the researchers were allowed to attend the recording session and take observational data during it.

3.3. Quality Assessment by Users

Users assess the quality through a user-reception study: they watch several final versions of the AVT product and answer questions in order to assesse whether the translation or post-editing meets its purpose. In the case of the study presented in this article, the purpose of the translated or post-edited product was infotaining – informative and entertaining at the same time. Therefore, the reception study evaluated both the users enjoyment to it and the level of understandability by the users.

4. Methodology

The experiment to assess the quality of wildlife documentaries to be voiced-over and off-screen dubbed was carried out with only one language combination: English into Spanish. The aim of the experiment is to assess the quality of post-edited wildlife documentaries compared to the quality of human translations and it is built upon the hypothesis that there is no significant difference between the quality of post-editing and the quality of translation of wildlife documentaries in English delivered through VO and OD in Spanish.

4.1. Participants

The raters taking part on the first level of QA of this study were six lecturers of MAs on audiovisual translation in universities in Spain who are, at the same time, experts on voice-over and currently work or have recently worked as professional voice-over translators. The raters' profiles are comparable: all of them have a BA in Translation Studies except for one, who has a BA in German Philology. Furthermore, five of them have either a PhD in Translation or have attended PhD courses on the same field. From the six raters, two are more experienced than the others: when the experiment was carried out raters 1, 3 and 5 had worked as audiovisual translators between 10 and 16 years and taught for 11, 8 and 5 years respectively, while participants 2, 4 and 6 had between 5 and 8 years of experience as audiovisual translators and taught for the last 4 or 5 years. The number of raters used is higher than the number of raters found in previous studies on QA and post-editing (Gerberof, 2009; García, 2011 or De Sutter et al, 2012)

For the second level of QA, however, only one dubbing study was used, as only one study participated in the recording of the materials.

For the third level of QA, 56 users participated on the study (28 men and 28 women). The ages of the participants ranged from 23 to 65 years old and from very different backgrounds. All participants were native speakers of Spanish and bout half of them were highly proficient in English.

4.2. Materials

The materials used for the first level of the experiment were 6 translations and 6 post-editings of two excerpts of a 7-minute wildlife documentary film titled *Must Watch: a Lioness Adopts a Baby Antelope* that is currently available on Youtube as an independent video (http://www.youtube.com/watch?v=mZw-1BfHFKM) although it is part of the episode *Odd Couples* from the series *Unlikely Animal Friends* by National Geographic broadcasted in 2009. The translations and post-editings (24 in total) were produced by 12 students of an MA on audiovisual translation that had had a specific course on voice-over. The excerpts are similar in terms of length, number of words and entries. They were machine translated through Google Translate MT engine, the best free online MT engine to be used to machine translate wildlife documentary scripts according to Ortiz-Boix (forthcoming). Automatic measures show that the machine translation of the first excerpt (BLEU: 44.97; TER: 69.17) was ten points better than the machine translation of the second (BLEU: 33.75; TER: 59.68).

For the second and third levels, only the best post-editing and the best translation of both excerpts, selected according to the quality assessment produced by the raters, were used. However, while the scripts were used for the second level, the final audiovisual product was used in the third one.

4.3. Test Development

The experiment was divided in three parts, which were done consecutively in the time. The first part corresponds to the first level (QA by raters), the second part to the second level (QA by dubbing studio), and the third part to the third level (QA by users).

1st part. Participants carried out the experiment from their usual place of work by following detailed instructions to assess the 24 documents without knowing which of them were translations or post-editings. They were given 20 days to perform the whole QA assessment. The experiment was divided in two rounds of evaluation: For the first round raters, who were given just one day, received the 24 documents in a different order. They were instructed to read each document and grade it according to their first impression on a 7-point scale (scoring round 1). For the second round, they were not given any specific time, as far as they delivered the QA within the 20 day they were given and they did not correct a same document in two different days. In this round of assessment, raters were asked to correct the documents following a specific evaluation matrix (see section 5.4) and grade the documents after the correction on a 7-point scale (scoring round 2). Afterwards, they were instructed to answer an online questionnaire on their opinion about the document they corrected (questionnaire-based evaluation, see section 3.4.), and give each document a final grade between 0 and 10 (scoring round 3). They were also asked to guess whether the assessed document was a translation or a post-editing (post-editing/translation identification task).

2nd part. This QA level was carried out in the dubbing studio. They were asked to modify the given text, if necessary, while they recorded. They were also asked to follow the same criteria they usually use when recording. The researchers took observational notes of every change that were annotated first and counted later in order to be able to assess the quality of the translated and post-edited scripts being recorded.

3rd part. The last QA level was carried out in a lab environment that recreated the conditions in which documentaries can be watched: they were sited in an armchair and watched the documentary in a 32' flat screen. Participants were shown 2 of the excerpts without knowing if they were watching a translated or post-edited excerpt. Before watching the first excerpt, participants answered a questionnaire on demographic aspects (e.g. age, studies) and audiovisual habits. After watching the first excerpt, they answered an enjoyment and a comprehension questionnaire. Afterwards, participants were shown the second excerpt and after it they answered the same questionnaires again. The enjoyment questionnaire had 12 questions, whereas the comprehension one had only 5, as it was designed only to assess the general comprehension of the text.

4.4. Evaluation Matrix and Questionnaire Design

The evaluation matrix used for the first level of this QA metric is based on MQM, as it is designed to assess human and machine translations, as well as post-editings. Furthermore, it allows checking only the relevant categories for the text type used for this study, as well as including domain specific issue types.

Although MQM offers the possibility to include over one hundred issue types, only five categories and eleven subcategories of issue types were selected (See Table 1).

Issue types categories	**Issue types subcategories**
Adequacy	Wrong Translation
	Omission
	Addition
	Non-translated words
Fluency	Register
	Style
	Inconsistencies
	Spelling
	Typography
	Grammar
	Others
Variety	
Voice-over specificities	Spotting
	Action synchrony
	Phonetic transcriptions
	Isochrony
Design/Layout	
Others	

Table 1. *Evaluation matrix: error typology*

The selection was based on previous research on errors produced by MT engines in general texts (Avramidis et al., 2014) and wildlife documentary films (Ortiz-Boix, forthcoming), as well as in post-editings (Guerberof Arenas, 2009). As MQM does not include a domain specific issue type for audiovisual translated texts, a new category has been added: VO/DO specificities, including issue types spotting, action synchrony and voice-over isochrony and inclusion of phonetic transcriptions. Raters assessed the documents following an explanatory document. As far as the questionnaire is concerned, participants had to report their level of agreement with eight statements to assess general fluency, grammar, spelling, vocabulary appropriateness, terminology coherence, voice-over specification standards, satisfaction for the final result and purpose meeting.

The evaluation matrix used for the third QA level is based on the reception study layout by Gambier (2009). Each used questionnaire contained open and multiple choice questions depending on its purpose: while the questionnaire on demographic aspects contained open questions, the questionnaire on the audiovisual habits contained multiple choice questions based on a 4 or 7-point likert scale. The enjoyment questionnaire, which tested issues related to likability, attention, enjoyment and interest, as well as questions to determine the perceived quality of the text (they were asked if they noticed errors in the target language), contained basically multiple choice questions. The comprehension questionnaire had questions to determine the perceived and the actual intelligibility of the text.

4.5. Data and Methods

For the first QA model level, forty-eight corrected documents were collected along with the corresponding questionnaires. For the second QA level of the model, the correction by the dubbing studio of the four selected documents to be voiced-over and off-screen dubbed were collected. Finally, for the third QA level, 56 responses for every questionnaire were collected. All the data was analyzed using the statistical system R-3.1.2, developed by John Chambers and colleagues at Bell Laboratories. The following data were obtained:

(1) *First QA level (raters):*

1. The grades for each document in the three scoring rounds.
2. 144 questionnaire (6x24 documents) reporting on the participants' opinions after correcting each document.
3. The results of the post-editing/translation identification task.
4. 144 documents with corrections (6x24) according to the evaluation matrix based on MQM.

(2) *Second QA level (dubbing studio)*

1. 4 documents with corrections (1x4) made by the dubbing studio.
2. Observational data on the changes and difficulties taken in the dubbing studio.

(3) *Third QA level (end users):*

1. 56 questionnaire responses reporting on demographic aspects and audiovisual habits.
2. 112 questionnaire responses (14x4) on users enjoyment and comprehension.

An ANOVA variance test was carried out to validate the results. Statistical significance is assumed for P>0.05, i.e. the difference between the results of the results for the post-editing and the translation QA should be higher than 0.05 to be considered significant.

5. Discussion of Results

As the QA was performed in three levels (QA by raters, QA by dubbing studio and QA by users), the discussion of results is presented accordingly.

5.1. Quality Assessment by raters[12]

The results of the QA by raters, which are presented in detail in Ortiz-Boix and Matamala (forthcoming), indicate that the quality of both the translations and the post-editings carried out by the MA students is rather low. However, they prove that there are no significant differences between post-editing and translating from scratch as far as quality is concerned.

A summary of the results is presented below: the holistic approach contains the scores the raters gave in three rounds, the answers to the questionnaires and the categorization of the documents. The analytic approach contains the results of the corrections performed by the raters.

[12] See Ortiz-Boix (forthcoming) for further information on the results of the QA by raters.

5.2. Holistic Approach: Scores

The results for the first round of evaluation indicate that raters evaluate better translations than post-editings after reading the documents for the first time: while 45 out of 72 translations were evaluated from "pass" to "excellent", only 37 out of 72 post-editings were evaluated within that range of scores. However, when raters grade the documents after correcting them (second round of evaluation), the difference between post-editing and translations are reduced. In this case, 41 out of 72 translation and 38 out of 72 post-editings are rated from "pass" to "excellent". All in all, the mean grade in both the first and the second rounds of evaluation is only "pass".

The results for the third round of evaluation, which is more specific as raters were given more options in order to evaluate the document, show that the difference between translating from scratch and post-editing is minimal, as the mean grade for translations is 5.44 out of 10, being 10 the highest grade, and 5.35 out of 10 for post-editings. Both grades correlate perfectly with the grades from the first and the second rounds, as the numerical equivalence for "pass" is a number between 5 and 6.

Thus, results seem to indicate that more than one round of scoring are necessary in order to better determine the quality of the documents, as they show whether the grades are consistent throughout the whole QA process.

5.3. Holistic Approach: Questionnaire

Raters assessed from 1 to 7, being 7 the highest mark, 8 issue types: style, grammar, spelling, vocabulary, terminological coherence, VO specificities, correction, dubbing studio. Results indicate that post-editings are given higher grades in 4 of the issue types – grammar, terminological coherence, correction and dubbing studio – and the exact same grade in the case of issue type VO specificities. Therefore, when the results of the questionnaire are correlated with the scoring rounds, it can be assumed that issue types grammar, terminological coherence, correction and dubbing studio are less valuable to raters than the others, as they were assessed better for post-editing in the questionnaire although post-editings received lower grades than translations.

5.4. Holistic Approach: Categorization

According to the results, evaluators categorized correctly 42 translations out of 72 and only 22 out of 72 post-editings. They categorized wrongly 14 translations out of 72 and 27 out of 72 post-editings and could not decide whether the document was a translation or a post-editing in the case of 16 out of 72 translations and 23 out of 72 post-editings. Thus, results indicate that post-editings are more difficult to identify than translations, as the great majority of them are either misidentified or not recognized as such, which might lead to the conclusion that the quality of both post-editings and translations is similar, as also indicate the results of the other levels of the holistic approach to QA.

5.5. Analytic Approach: Correction

According to the corrections performed by the raters, translations need a lower number of corrections than post-editings. However, the difference between the number of corrections is not high, as the largest difference between post-editings and translations is of 5 corrections. According to the results, raters did not correct any error regarding synchrony and did a higher number of corrections for post-editings in all issue types but three: omission, addition and spelling.

All in all, results seem to prove that this level of the QA model is valid in case of VO and DO transfer modes, as the results of each part correlate with the others. However, it should also be tested with other transfer modes changing the specifications regarding VO and DO for those involved in the transfer mode to be tested.

6. Quality Assessment by Dubbing Studio

The dubbing studio only worked with the best translation and the best post-editing of each excerpt, according to QA by raters. In order to assess the quality of the translations and post-editings, the dubbing studio took basically only three things into consideration: synchrony, repetitions and phonetics. Thus, the results of the QA by the dubbing studio indicate that in general there is not much difference between the translations and the post-editings carried out by the MA students.

However, although no significant difference between the quality of the post-eiting (6 changes were made regarding synchrony and phonetics) and the translation (5 changes were made regarding synchrony and phonetics) are found in the first excerpt, the difference is noticeable in the case of the second excerpt. In the case of the second excerpt, only 4 changes regarding synchrony and phonetics were made in the translated version but no change was made in the case of the post-editing. According to the studio, the quality of the post-editing in terms of synchrony was so low that the only way to improve its quality was to retranslate it. Finally, it was decided to record the post-edited version although the VO isochrony could be affected.

7. Quality Assessment by Users

The results of the QA by users indicate that there is no significant difference between the quality of the translations and the post-editings carried out by the MA students. As users were divided in two age groups – -40 years old (group A) and +40 years old (group B) –, results are presented accordingly. The users within group A have a mean age of 25.45 years old and the users within group B, a mean age of 53 years old. The results on demographic issues indicate that the consumer habits do not vary much between group A and B. Although 54 out of 56 participants watch a maximum of 3 documentary films on TV every month, only 8 out of the 56 claim to watch between two and three wildlife documentary films every month. However, when they were asked what transfer mode they prefer (subtitling, dubbing or VO) when watching documentaries, the groups do not agree. While group A prefers watching subtitled (50%) or dubbed (46.43%) documentaries, group B prefers them voiced-over (46.43%) or dubbed (42.86%). The great majority of the users of both age groups claim that they prefer watching other programs such as reality shows or cooking programs dubbed (group B: 82.14%; group A: 60.71%) than voiced-over (14.29% for both groups).

The results show, on the one hand, that there is no significant difference between post-edited and translated wildlife documentary films (p=0.945) for none of the age groups regarding the enjoyment questionnaire (further results in section 6.3.1). On the other, the results also show no significance difference between post-edited and translated wildlife documentary films (p=0.864) for none of the age groups in regards of the comprehension questionnaire (further results in section 6.3.2).

7.1. Enjoyment Questionnaire

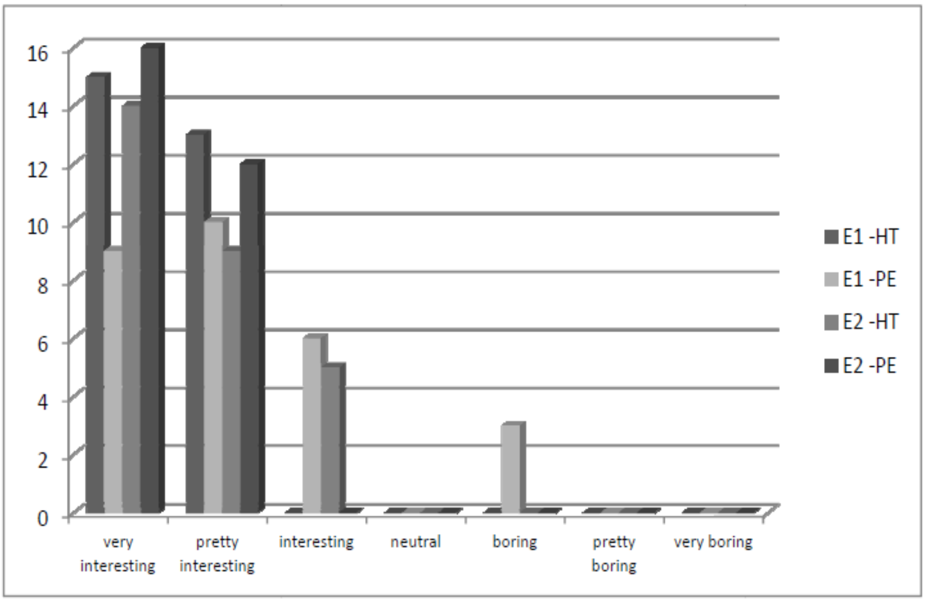


Figure 1. *Interest*

According to the results regarding the first excerpt, both age groups grade as more interesting the translated version than the post-edited one (see Figure 1), which was even qualified as boring by 3 users. However, the difference is minimal, as the median in both cases indicates that the excerpt is "pretty interesting". At the same time, they enjoyed watching the translated version more than the post-edited (see Figure 2). Results also indicate that users followed the excerpt actively, they were focused on what they were watching on screen and they lost the track of time. When users were asked on specific issues related to VO transfer mode, 24 out of 28 of the users in the case of the post-edited version, and 28 out of 28 for the translated, indicated that the combination of voices did not bother them.

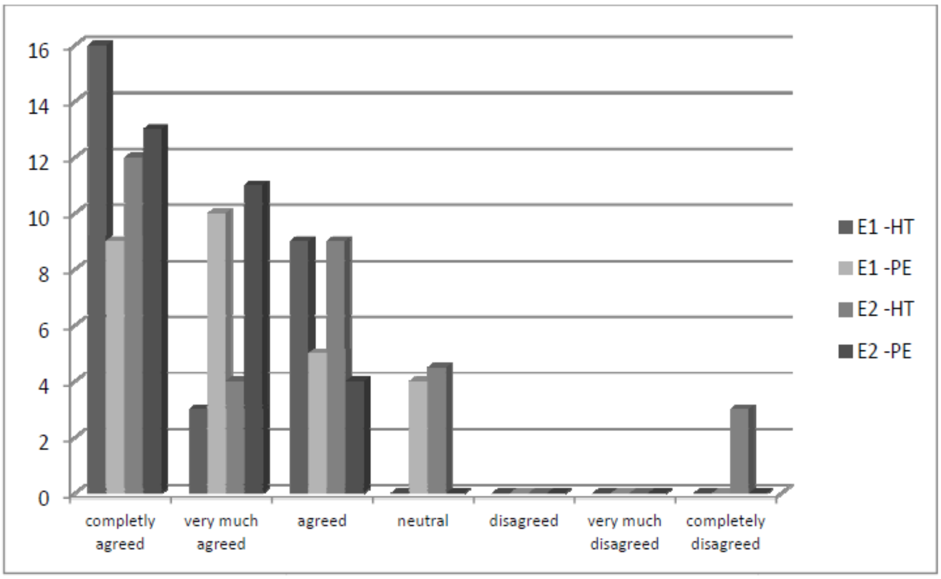


Figure 2. *Enjoyment.*

The results regarding the second excerpt, however, show that the post-edited version is rated higher than the translated one (see Figure 1), even though the quality of this post-edited version was rated poorly in terms of quality by the dubbing studio. The difference is wider for this excerpt, as the median in the case of the post-edited version indicates the post-edited excerpt is "very interesting" while the translated version is "pretty interesting". Nevertheless, users enjoyed both versions equally (Figure 2). In this excerpt's case, results also indicate that users followed the excerpt actively, they were focused on what they were watching on screen and they lost the track of time. Finally, when they were asked on specific issues related to VO transfer mode, 22 out of 28 users who watched the post-edited version and 20 out of the 28 who watched the translated one indicated the use of two voices did not bother them.

In conclusion, the results show that the first customer requirement (enjoyment) has been fulfilled in the case of both excerpts and, thus, their quality could be labeled as "good". The results also indicate that there is no significant difference between the post-edited and translated versions according to the users point of view.

7.2. Comprehension Questionnaire[13]

Before answering the comprehension questions, users were asked on their perceived comprehensibility of the excerpt. All the users form both age groups (56 out of 56) agreed on the fact that both excerpts, independently from the version they watched, was completely comprehensive and it contained no expressive nor linguistic problems. However, the comprehension questionnaire shows otherwise, as the comprehension level was not as high as they claimed it was.

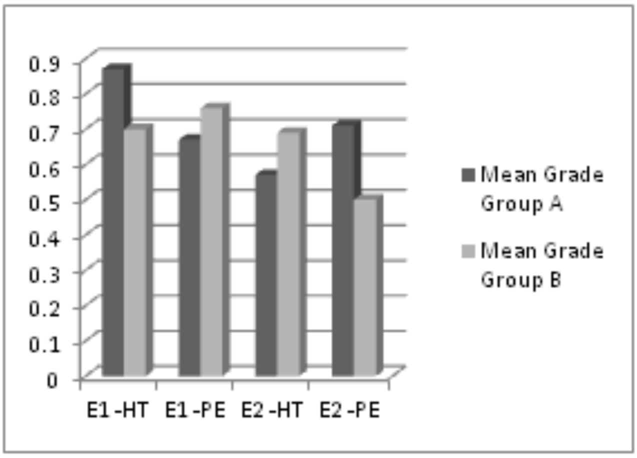


Figure 3. *Grades.*

The results show that the difference on the comprehension between the translated and the post-edited version of the first excerpt is minimal for the age group B (see Figure 3). However, the difference increases in favor of translation for the age group A. In the case of the second excerpt, the difference is of more than 0.15 points in both cases. Nevertheless, the difference is in favor of the translated version for the age group B and in favor of the post-edited version for age group A. All in all, the comprehension is above 0.5 in all of the cases, suggesting that the comprehension level is not low.

[13] In order to analyse this questionnaire, the researchers gave a grade to each answer depending on whether the answer was good (1 point), more or less good (0.5 points) or wrong (0 points).

In conclusion, as the results show that the comprehension is good for both excerpts, the second consumer requirement (informative) is also achieved. Therefore, the results indicate that the product is of quality according to the end-users. The results also indicate that the quality of post-edited version is similar to the quality of the translated one.

8. Conclusion and Further Research

This article presents a QA proposal to evaluate post-edited wildlife documentary films in comparison to translated documentaries. The QA was carried out in three levels, including user reception as well as an assessment by a dubbing studio to the usual QA performed by human raters. The results of the study indicate that there are no significant differences between translated and post-edited wildlife documentary films when working on the language combination English into Spanish. However, when the three levels of QA are compared, it can be inferred that the expectations of the users are not high, as they rated high all the versions of the excerpts even though the human assessors rated them of rather low quality and the dubbing studio even said one of the post-edited versions should have been retranslated. The low quality of both the translations and post-editings might be due to the fact that the translators used for the experiment were MA students. Thus, it remains to be seen whether the presented results would be corroborated if professional audiovisual translators translated and post-edited the excerpts, as their resulting translations and post-editings would be assumed to be of greater quality.

References

Bywood, L., Volk, M., Fishel, M., and Georgakopoulou, P. (2013). Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives, 21*(4), 595-610.

Carl, M. Dragsted, B. Elming, J. Hardt, D. Jakobsen, A. 2011. "The process of postediting: a pilot study". In *Proceedings of the 8th international NLPSC workshop.* Bernadette

Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds). (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur: 131-142.

Chesterman, Andrew (2007): 'Bridge Concepts in Translation Studies'. In Wolf, Michaela and Fukari, Alexandra (eds): Translation Sociology. A new discipline under construction. Selected papers from the International conference on Translating and Interpreting as a Social Practice, Graz, 5-7 May 2005. Amsterdam and Philadephia: John Benjamins.

Etchegoyhen, T., Fishel, M., Jiang, J., and Maucec, M. S. (2013). SMT Approaches for Commercial Translation of Subtitles. In *Machine Translation Summit XIV, Main Conference Proceedings* (pp. 369-370).

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

Franco, E., Matamala, A., and Orero, P. (2010). *Voice-over translation: An overview*. Peter Lang.

Gambier, Y. (1998). *Translating for the Media*. University of Turku.

Gambier, Y. (2008). "Recent developments and challenges." *Between text and image: Updating research in screen translation* 78: 11.

García, I. 2011. “Translating by post-editing: Is it the way forward?” *MachineTranslation*, Vol. 25(3). Netherlands: Springer. 217-237

Guerberof Arenas, A. (2009). "Productivity and quality in MT post-editing." *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT.*

Guerberof Arenas, A. (2012). *Productivity and quality in the post-editing of utputs from translation memories and machine translation*. . Universitat Rovira i Virgili. Mariana, 2014

Hansen, Gyde. "The speck in your brother’s eye–the beam in your own." *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile* 80 (2008): 255.Melby, 2014

House, J. (2006). Text and context in translation. *Journal of Pragmatics, 38*(3), 338-358.

Kovačič, Irena (1995): 'Reception of subtitles. The non-existent ideal viewer'. In Translatio (FIT Newsletter) 14 (3-4). 376-383.

Matamala, A. (2009) "Main Challenges in the Translation of Documentaries." *New Trends in Audiovisual Translation*. Ed. Díaz Cintas, Jorge. Bristol: Multilingual Matters. Chapter 8.: 109-120

Melby, A., Fields, P., Koby, G. S., Lommel, A., and Hague, D. R. (2014). Defining the Landscape of Translation. In *Tradumàtica* (pp. 0392-403).

Melero, Melero, et al. (2006). "Automatic Multilingual Subtitling in the eTITLE project." *Proceedings of ASLIB Translating and the Computer 28*. London.

Nida, E. A. (1964). *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Orero, P. (2006) Voice-over: A case of hyper-reality. *EU High Level Scientific Conference Series. MUTRA Proceedings*.

Ortiz-Boix, Carla. (Forthcoming). "Post-Editing Wildlife Documentaries: Challenges and Possible Solutions."

Ortiz-Boix, C. and Matamala, A. (forthcoming). Post-Editing Wildlife Documentary Films: a new possible scenario?

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

Pönniö, Kaarina. "Voice over, narration et commentaire." *Communication audiovisuelle et transferts linguistiques/Audiovisual communication and language transfers*. Ed. Gambier, Yves. International Forum, Strasbourg, 22-24 June 1995. Special issue of *Translatio* (FIT Newsletter / Nouvelles de la FIT). (1995): 303-307.

Reiß, K., and Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie* (Vol. 147). Walter de Gruyter.

Chapter 6. Summary

6. Summary

6.1. Summary in English

This PhD explores the possibility of introducing MT and PE MT into the process of translating wildlife documentaries to be voiced-over and off-screen dubbed. In order to do so, 3 studies have been carried out:

The first study, which is presented in the first article included in this dissertation, researches the challenges that would need to be overcome if MT is to be included in the process of translating wildlife documentaries through VO and OD. The challenges are based on previous studies on the translation via VO and OD, and the translation of documentaries (e.g. Franco et al., 2010), as well as an analysis of a corpus of wildlife documentary scripts in English, a corpus of wildlife documentary scripts in Spanish, and a corpus of segments of documentary scripts and 9 translations: 8 produced by free-online MT engines and 1 by a human translator. Furthermore, it presents possible solutions to such challenges, which are used in the experiment presented in the second study.

The second study, which represents the second article presented in this PhD, presents and experiment that intends to determine the effort required to post-edit wildlife documentary films, as compared to the effort needed to translate them from scratch. 12 MA students participated in the experiment, which was based in Krings (2001) notion of PE effort, and previous experiments on PE effort (e.g. O'Brien, 2006). The conclusions of the experiments seem to indicate that PE requires less effort than translating them from scratch.

The third study, which is presented in the third and fourth articles of this dissertation, compares the quality of post-edited and translated wildlife documentaries. In order to do so, two experiments were carried out. The first one presents the QA, made by 6 experts, of the documentary scripts translated by MA students during the second study. The experts assessed the quality by grading the documents, by correcting them using an MQM-based error classification that includes specifications of documentary translation, and finally, by answering questionnaires on their opinions. Conclusions indicate that the quality of post-edited and translated wildlife documentaries is significantly similar.

In the second experiment of the third study, 56 end-users blindly assessed the quality of post-edited and translated wildlife documentaries by using user-reception questionnaires. As well as the QA by experts, the results show that there is no significant difference between the quality of post-edited and translated documentaries.

All the studies presented in this dissertation help to accomplish its main objective: "research whether MT might be successfully included, effort and quality wise, into the process of translating documentaries of a certain subdomain (wildlife) through VO and OD"; and validate its main hypotheses, which stand that "the inclusion of MT into the process of translating wildlife documentaries through VO and OD will optimize the process in terms of effort" and that "the inclusion of MT into the process of translating wildlife documentaries through VO and OD will not have significant impact on the quality of the translated product".

6.2. Summary in Spanish

Esta tesis doctoral explora la posibilidad de incluir traducción automática y traducción automática seguida de post-edición en el proceso de traducción de documentales científicos de fauna y flora mediante voces superpuestas y doblaje en *off*. Para conseguirlo, llevamos a cabo tres estudios:

El primer estudio, que se incluye en el primer artículo de la tesis, investiga los desafíos que necesitamos superar para poder incluir la traducción automática en el proceso de traducción de documentales. Nos basamos en estudios anteriores sobre traducción mediante voces superpuestas y doblaje en *off* y sobre traducción de documentales (por ejemplo, Franco et al., 2010) para determinar los desafíos que nos podemos encontrar. También nos basamos en el análisis de un corpus de guiones de documentales en inglés, un corpus de documentales en español y un corpus de segmentos de guiones de documentales y sus nueve traducciones, ocho de las cuales fueron producidas por motores de traducción automática gratuitos y en línea, y otra traducción hecha por un traductor humano. Además, el estudio introduce posibles soluciones para estos desafíos, que usamos en el experimento presentado en el segundo estudio.

El segundo estudio, que se incluye en el segundo artículo de la tesis, se centra en un experimento que pretende determinar el esfuerzo requerido para post-editar documentales de naturaleza comparado con el esfuerzo necesario para traducirlos. Doce estudiantes de máster participaron en el experimento, que se basó en la noción de esfuerzo de Krings (2001), así como otros estudios relacionados con esfuerzo de post-edición (por ejemplo, O'Brien, 2006). Las conclusiones del experimento parecen indicar que post-editar requiere menos esfuerzo que traducir.

El tercer estudio, que se encuentra en el tercer y cuarto artículos de la tesis, compara la calidad de documentales post-editados y traducidos. Para hacerlo, se llevaron a cabo dos experimentos. En el primero, seis expertos evaluaron la calidad de los documentales hechos por los estudiantes durante el experimento: los expertos pusieron nota a los documentos, los corrigieron usando una clasificación de errores basada en el MQM que incluye especificaciones para la traducción de documentales, y respondieron unos cuestionarios sobre su opinión de las traducciones. Las conclusiones del experimento indican que la calidad de las post-ediciones y las traducciones son significativamente similares.

En el segundo experimento, 56 posibles usuarios evaluaron la calidad de documentales post-editados y de documentales traducidos mediante un cuestionario de recepción. Igual como indicó la evaluación de los expertos, los resultados muestran que no hay diferencias significativas entre la calidad de los documentales traducidos y los post-editados.

Los estudios presentados en esta tesis doctoral, pues, nos permiten conseguir el objetivo principal de la tesis: "investigar si se podría incluir la traducción automática de manera satisfactoria, en cuanto a esfuerzo y calidad, en el proceso de traducción de documentales de un subdominio específico (fauna y flora) mediante voces superpuestas y doblaje en *off*"; y validar las dos hipótesis principales de la tesis: "la inclusión de traducción automática en el proceso de traducción de documentales científicos de fauna y flora mediante voces superpuestas y doblaje en *off* optimizará el proceso en cuanto a esfuerzo" y "la inclusión de traducción automática en el proceso de traducción de documentales científicos de fauna y flora mediante voces superpuestas y doblaje en *off* no afectará significativamente la calidad del producto traducido".

6.3. Summary in Catalan

Aquesta tesi doctoral explora la possibilitat d'incloure la traducció automàtica i la traducció automàtica seguida de postedició en el procés de traducció de documentals científics de fauna i flora mitjançant veus superposades i doblatge en *off*. Per a aconseguir-ho, es van portar a terme tres estudis:

El primer estudi, que s'inclou en el primer article de la tesi, investiga els reptes que hauríem de superar per a poder introduir la traducció automàtica en el procés de traducció de documentals. Ens basem en estudis anteriors sobre traducció mitjançant veus superposades i doblatge en *off* i sobre traducció de documentals (per exemple, Franco et al., 2010) per a determinar els esmentats reptes, així com en l'anàlisi d'un corpus de guions de documentals en anglès, un corpus de documentals en castellà i un corpus de segments de guions de documentals i 9 traduccions, 8 de les quals van ser fetes per diversos motors de traducció automàtica en línia i gratuïts, i l'altra va ser traduïda per un professional. A més, l'estudi introdueix possibles solucions als reptes, que fem servir a l'experiment que presentem en el segon estudi.

El segon estudi, que s'inclou en el segon article de la tesi, se centra en un experiment per a determinar l'esforç necessari per a posteditar documentals científics de fauna i flora en comparació amb l'esforç requerit per a traduir-los. Dotze estudiants de màster van participar en l'experiment, que es va basar en la noció d'esforç de Krings (2001), així com en altres estudis sobre esforç de postedició (per exemple, O'Brien, 2006). Les conclusions de l'experiment indiquen que post-editar requereix menys esforç que traduir.

El tercer estudi, que es pot trobar en el tercer i el quart articles de la tesi, compara la qualitat de documentals post-editats i traduïts. Per a fer-ho, es van portar a terme dos experiments. En el primer, sis experts van avaluar la qualitat dels documentals que van fer els estudiants de màster durant l'experiment del segon estudi. Així, doncs, els experts van posar nota als documents, els van corregir mitjançant una classificació d'errors basada en MQM que inclou especificacions per a la traducció documentals, i van respondre qüestionaris d'acord amb la seva opinió sobre les traduccions corregides. Les conclusions de l'experiment indiquen que la qualitat de les post-edicions i de les traduccions són significativament semblants.

En el segon experiment, 56 possibles usuaris van avaluar la qualitat dels documentals post-editats i traduïts mitjançant un qüestionari de recepció. Igual com va indicar l'avaluació dels experts, els resultats mostra que no hi ha diferencies significatives entre la qualitat dels documentals traduïts i els posteditats.

Els estudis presentats en aquesta tesi doctoral, doncs, permeten aconseguir l'objectiu principal de la tesi: "investigar si es podria incloure, satisfactòriament, la traducció automàtica, en quant a esforç i qualitat, en el procés de traducció de documentals d'un subdomini específic (fauna i flora) mitjançant veus superposades i doblatge en *off*; i validar les dues hipòtesis principals de la tesi: "esperem que la inclusió de traducció automàtica en el procés de traducció de documentals de naturalesa mitjançant veus superposades i doblatge en off optimitzi el procés en quant a esforç" i "esperem que la inclusió de traducció automàtica en el procés de traducció de documentals científics de fauna i flora per veus superposades i doblatge en *off* no afecti significativament la qualitat del producte traduït”.

Chapter 7. Conclusions

7. Conclusions

This PhD started with the aim of studying the possible implementation of MT into VO and OD, two oral AVT transfer modes. As VO and OD are mostly applied to factual programs, it was decided to research the inclusion of MT into the process of translating documentaries of the wildlife subdomain to be voiced-over and off-screen dubbed. Hence, the main objective of this research study was to research whether Machine Translation might be successfully included into the process of translating documentaries of a certain subdomain through VO and OD, focusing on the effort involved in such process and the quality of the resulting translation.

In order to reach this main objective, three secondary objectives were set: (1) Determine what characteristics of documentary translation through VO and OD would be challenging when including MT into the process and propose solutions to overcome such challenges; (2) Compare the effort required to post-edit machine translated documentaries to be voiced-over and off-screen dubbed to the effort of translating them from scratch; and (3) Assess the quality of post-edited documentaries to be voiced-over and off-screen dubbed as compared to the quality of translated documentaries.

This last chapter will begin by determining whether the secondary objectives have been accomplished. This will lead to a discussion on the accomplishment of the main objective of this PhD, while validating and/or refuting the hypotheses presented in the introduction. Likewise, I will go over the more important contributions of this research study, and dive into the implications of the results for the AVT industry in the future.

The chapter will be divided into two sections: the first, Discussion on the Results, will discover if the secondary objectives presented in the introduction were accomplished and if the hypotheses related to each of them were validated or refuted. Furthermore, the main objective and hypothesis will also be addressed. In the second section, the main contributions of this dissertation, as well as its limitations, and possible follow-up studies for this research will be presented and discussed.

7.1. Discussion on the Results

As this PhD has a main objective and three secondary objectives, this section will be divided into four subsections: in the first, the results and conclusions of the challenges for MT to be introduced into documentary translation will be discussed (1st article). In the second, the outcome of the study on PE effort as compared to the effort of translation from scratch (2nd article) will be conferred. The third subsection will discuss the results and the conclusions of the quality assessments (3rd and 4th articles). Finally, in the last subsection the results will be wrapped up and the overall results and the main hypothesis will be discussed.

7.1.1. *On the Specificities of Translating Documentaries and their Impact on Machine Translation*

The first secondary objective of this research was descriptive; it aimed to set the characteristics of the translation of documentaries through VO and OD from a translator's point of view in order to determine which ones might be challenging if MT was included into this particular translation process. Furthermore, it also worked towards proposing ways to help overcome the challenges and successfully implement MT into VO and OD AVT modalities and set the theoretical framework of this PhD. In order to achieve this objective, I first explored previous research studies on VO and on the translation of documentaries. Afterwards, three corpora of documentaries in English and documentaries in Spanish were created, analyzed and compared. Finally, machine-translated segments of documentary scripts were assessed. The analyses allowed us to identify 9 main challenges, which were divided in three groups, depending on the procedures used to distinguish them.

After the review of the research studies on VO and translation of documentaries, three fundamental challenges were discerned:

(1) *Spotting*. Even though spotting is essential for any audiovisual product to be recorded, audiovisual translators, who oftentimes work either with original scripts that do not include time codes or with original scripts that include wrong spotting, have to either modify the provided time codes or include them (Franco et al., 2010). Hence, spotting would be as much as a challenge for MT as it is for audiovisual translators nowadays. In order to reduce PE effort and increase productivity, it is recommended to introduce, check and/or correct the time codes in the original script before feeding it into the MT engine, as post-editors would not have to deal with the tedious work of spotting the script while they revise and correct the MT output. For the experiment on post-editing effort, thus, participants were given a formatted script with no problems regarding spotting. However, they were asked to revise them and change the spotting if they considered it was erroneous.

(2) *Synchronization* is a key feature in audiovisual translation in general. Documentary translation through voice-over and off-screen dubbing has to deal with three types of synchronies: kinetic and action synchronies, and VO isochrony (Orero, 2006). Synchronization could not be reached without the essential work of translators, who rephrase, condense, or adapt the text in order to make the audiovisuals match in the available time slots. Therefore, if MT is to be introduced in documentary translation, post-editors should not only revise and correct the text linguistically, but also in terms of synchronization. Therefore, they would need a piece of software that allowed them to work with audiovisuals, a problematic that is addressed in the next point. Furthermore, automatic strategies such as quality estimation could be implemented in order to reduce the workload. In the first experiment presented in this thesis participants

were asked to post-edit the excerpts taking into account the synchronization constraints.

(3) *Access to the audiovisual content.* In AVT the source text is not a written document such as a script, but the audiovisual product per se (Espasa, 2006; Franco et al., 2010). However, MT engines still need a written document in order to produce a translation and do not take yet audiovisuals into account, that is why PE has been only successfully implemented in subtitling, an AVT written modality (Etchegoyhen et al., 2014). Throughout the first article it has been proven that post-editors are not only recommended to check the translated scripts in terms of language adequacy and fluency, as usual, but a strong emphasis has been put on the revision of the correlation between written text and audiovisual content, as well as synchrony. In order to do so, post-editors would need access to the audiovisuals, which, to the best of my knowledge, is not yet included in any post-editing software. Then, post-editing software that has the possibility of integrating visuals and audio is currently a great challenge to be overcome if MT is to be included not only into the translation of documentaries but also into other audiovisual translation modalities.

After the analyses of documentary scripts in English and Spanish, five challenges were pinpointed:

(1) *Variety of the script format.* Currently, there is no standardization for the layouts of the original or translated scripts. However, not only is the huge variety of scripts problematic, but also the amount of information some original scripts include: translations only include time codes, name of the talking heads or narrator, and translation of their words. Thus, in order to include MT into the process, an adapted translation script or template without extra information should be created before feeding it to the MT engine. In order to do so, a pre-

editing phase, where all the extra information is deleted and the essential information is added into a basic template, is advised and proposed as a solution. Analyzing pre-editing was beyond the scope of this PhD and therefore, no studies on this direction have been carried out. However, the participants on the experiments worked on a template only with the information necessary to translate.

(2) *Variety of registers within the same script*. As VO is used to translate the words of the experts and/or spontaneous dialogues on camera, and OD is mostly used for narrators off-camera, different speakers can coexist within a same documentary. Depending on the speaker and the communicative situation, the register can vary, making it possible for a same script to have different registers. To guarantee high quality, MT is usually used with texts with one register, mostly standard. Hence, the combination of registers within a script might be more demanding, as there might be interference between them. Furthermore, specific features of oral discourse may be difficult to fix automatically, as such features are usually erased by the translator in order to reach VO synchrony (Orero, 2006). The variety of registers and the features of spontaneous oral discourse might be, thus, a great challenge to overcome if MT is to be included into the process. To overcome this challenge, this thesis has proposed to create a domain specific engine that could be customized according the specific necessities of the documentary to be translated. However, even though it is believed that a customized engine would increase the quality of the raw MT output, as has been proven in studies on domain-specific engines applied to other translation modalities, register problems would not be completely solved. Hence, the post-editor is considered to be necessary, regardless of the quality of the output produced by the MT engine used to translate, in order not to have register problems in the final version.

(3) *Terminology* is a relevant feature of documentary films and it varies depending on the topic of documentary. Furthermore, even if dealing with the same general topic, every subfield has its specific terminology while sometimes coexisting with terminology of other fields. The richness of the terminology within every documentary is challenging for the MT engines, as they would have to be specifically trained for each documentary. Departing from what has been researched for domain specific texts (e.g. Wu et al., 2008), it seems that to create domain specific MT engines for each field, or even subfield, in order to avoid terminology problems. However, another problem might arise, as many different engines should be created. Hence, it is proposed to create a basic domain specific engine for each field that could be modified according to the necessities of each particular documentary. All in all, further research is needed in this topic, as it has not been researched in this PhD because of time constraints.

(4) *Errors and inaccuracies in the original script*. Original scripts can contain wrong dates, names of places or terminology, parts of the text can be missing or can be in the wrong place. Although such errors and inaccuracies would not affect the MT output, they could slow down the post-editing process. In order to overcome this challenge, scripts could be checked before being fed into the MT engine during a pre-editing phase, which could help reduce the effort needed to post-edited the documentary film. However, further research on this subject matter is needed, as the addition of a pre-editing phase might make the effort of the overall process increase and, hence, translation from scratch might be recommendable effort wise.

(5) *Linguistic inconsistencies in the original script*. Apart form errors and inaccuracies, original scripts commonly contain spelling, grammar

> and punctuation mistakes. Sometimes, the mistakes are due to the script layout, which might cut the sentences into incoherent syntagma. These inconsistencies might affect the efficiency of the MT engine and have an impact on the performance of the post-editors. In order to overcome this challenge, scripts could be checked before being fed into the MT engine while reviewing possible errors and inaccuracies. However, the problematic presented in the previous point could appear and, hence, further research on the possibility of including pre-editing into the process is needed.

The last challenge was addressed after analyzing the corpus of translations produced by 8 free online MT engines. The MT output was assessed automatically using BLEU and TER automatic measures first, and subjectively assessed by the researcher afterwards. According to both the automatic and the human assessment, the MT engine that performed better was Google Translate, which was afterwards selected as the MT engine to translate the script for the other experiments that are part of this PhD.

After analyzing the translations, it could be pinpointed that the most common errors produced by the MT engines when translating wildlife documentaries are related to agreement (186 out of 859), mistranslations (133 out of 859), and words left untranslated (86 out of 859). This type of errors could be fixed using hybrid domain specific engines containing glossaries and domain specific data to fix mistranslations and untranslated words, and grammar rules to fix agreement inconsistencies.

The analysis of the challenges led to three solutions to increase the efficiency of the MT engines and the process of PE. The first proposed solution advocates for the inclusion of pre-editing into the process, as it could be the answer to challenges such as spotting, oral features, linguistic mistakes in the original script, and errors and inaccuracies within the original script. Furthermore, it could allow faster post-editing, aspect that has been already proven by e.g. Sertan et al. (2014) in other contexts. Further research is yet needed in order to determine if it is feasible effort wise, as the pre-editing phase has not been included in the experiment on effort *per se*. A pre-editing phase has been however carried out by the researcher in order to avoid any of the described problems during the post-editing phase in the experiment. The second solution proposes the building of a domain specific engine to help overcome terminology problems and solve the more common mistakes produced by the engines. Finally, the third solution advocates for the inclusion of templates into the workflow in order to deal with the coexistence of different scripts, which is a still existing problematic within the field of AVT and could also help standardization. The analyses have also pointed out the importance of the audiovisual content and the need to include it into post-editing software.

All in all, the foremost secondary objective of this PhD has been successfully fulfilled. Its accomplishment allowed to determine the most important features to take into account if MT and PE were to be included into the process of translating wildlife documentary films to be voiced-over and off-screen dubbing, while setting a framework for MT of documentaries to be voiced-over and off-screen dubbed. Therefore, the analysis of these features allowed us to set the basis of this research and present recommendations for the translation of documentaries and the translation through VO and OD using MT and PE MT. Furthermore, I have proposed solutions to the challenges of including MT into the process of translation through VO and OD that would need to be further researched in the future. The solutions proposed to overcome the challenges were partially used for the experiments, as it was considered an scenario were a domain specific MT engine could not be built, and hence, Google Translate was used to translate the excerpts that would be used throughout the experiments that form this PhD.

7.1.2. *On Post-Editing Effort*

The second secondary objective was exploratory. It aimed to compare the level of effort required to post-edit machine translated documentary scripts to be voiced-over and off-screen dubbed to the level of effort needed to translate them from scratch. It was built upon the hypothesis that post-editing effort is lower than the effort of translating from scratch when translating wildlife documentary films.

To accomplish this objective, it was set an experiment in which 12 MA students both translated and post-edited two excerpts of a documentary. The experiment on effort was based on Krings' (2001) types of effort and O'Brien's (2006) and Lacruz et al.'s (2014) studies on how to measure such efforts. These studies have proved to be trustful and easy-to-use measures with keylogging software to determine technical and temporal efforts, in the case of O'Brien (2006), and cognitive effort in Lacruz et al.'s (2014) case. Using the keylogging software Inputlog, it was recorded and analysed the amount of time participants needed to perform each task (temporal effort), the keylogs and the mouse clicks and movements they did (technical effort), and the number of pauses done and their duration in order to calculate the effort in seconds (cognitive effort).

The general results of the experiment, which considered the effort for two excerpts, determined that post-editing wildlife documentaries requires less effort than translating them from scratch. However, when results of each excerpt were analyzed separately, they differed in the case of the second excerpt. The results for each type of effort showed that (1) post-editing requires less temporal effort, although results were only significant in the case of the first excerpt; (2) post-editing requires less technical effort, although results were again only significant in the case of the first excerpt; and (3) post-editing is less cognitively demanding, although results were yet again not significant in the case of the second excerpt.

The diverging results may be due to uneven technical skills of the participants, and/or the low number of participants in the experiment. All in all, further research is needed, as it was believed to be just a first step to determine whether the inclusion of PE MT in the process of translating documentaries through VO and OD is possible.

Apart from a global presentation of the results, divided according to excerpt and type of effort, the results were further analyzed and data was presented in an innovative way. It was analyzed in what part of the process participants devoted more effort: while working on the document and performing the task *per* se, while documenting themselves, or while checking the audiovisuals. The results in this matter suggest that, regardless of the excerpt, the effort concentrated in the main document while performing the task. Furthermore, results show that the difference between PE and translation effort increase significantly when only the effort devoted to the task *per se* is taken into account. However, no significant differences were found between PE and translation effort in the case of documentation effort, term proposed to name the effort devoted to use search engines or audiovisual display. Thus, the results indicate that it may be possible to use PE MT into the process of translating wildlife documentary films to be voiced-over and off-screen dubbed, subject to further research with professional translators, domain-specific MT engines and a pre-editing phase.

All things considered, it can be said that the second secondary objective has been accomplished, as the level of effort required to post-edit and to translate have been compared and analyzed. The analysis has allowed us to present three of the main contributions of this PhD: 1. study the possible inclusion of PE MT into the process of translating through an oral AVT transfer mode; 2. determine that post-editing requires less effort than translating from scratch also in the case of an oral AVT transfer mode as, up until now, PE effort had only been researched for written texts; and 3. divide effort according to the part of the translation process it is being undertaken. While accomplishing this objective, the results of the carried out experiment proved the null-hypothesis of the study and, hence, allowed us to go a step further in the fulfilling of the main hypothesis of this PhD, as it was proven that the level of effort decreases when wild life documentaries are post-edited instead of translated.

7.1.3. *On Quality Assessment*

The third secondary objective was also exploratory, as it aims to assess the quality of post-edited documentaries to be voiced-over and off-screen dubbed as compared to the quality of translated documentaries. In order to accomplish this objective, and thus validate its null-hypothesis, it was needed to prove that the quality of the post-editings was significantly equal to the quality of the translations. Hence, two experiments were designed. In the first one, the quality of the translations and post-editings produced during the experiment on effort by 12 MA students was blind assessed by 6 experts: professional audiovisual translators that are or have been lecturers at a Spanish university. The assessment was done using a mixed approach based on MQM (Lommel et al., 2014). In the second experiment, 56 possible users blind assessed two Spanish voiced-over versions of each excerpts. The versions selected for this experiment were the translation and the post-editing of each excerpt with the best mean results in the assessment carried out by the experts.

The first experiment assessed the quality using a mixed approach using three scoring rounds: questionnaire-based evaluations, post-editing and translation identification task, and an evaluation matrix-based assessment. The general results of the assessment, as well as the results of each evaluation system separately, for each excerpt indicate that there is no significant differences between the quality of post-edited and the quality of translated wildlife documentary films. Despite the non-significance of the overall results, the qualitative analysis shows that the quality of the translations is higher than the quality of the post-editing. Nonetheless, the differences between translated and post-edited texts vary depending on the excerpt, which might be because of the particularities of each of them.

According to the assessment through questionnaires, post-editing is generally assessed higher for terminology coherence and domain specific issues, while translation is rated more positively for fluency and general vocabulary. The assessment through an evaluation matrix indicates that the types of errors when translating and post-editing are different. As in line with the results presented in the first article of this PhD, post-edited versions contained more errors in terms of style and grammar, and translations had more errors regarding mistranslated words. The post-editing and translation identification task showed that experts identify more easily the translations than the post-edited documents, compelling us to claim that the difficulty to identify the post-edited versions indicates that their quality was similar to the quality of the translations.

The second experiment presented one of the main contributions of this PhD. It assessed the quality according to end-users, who evaluated the excerpts according to user reception questionnaires, and comprehension. The results of the analysis of the user reception questionnaires show better results for translation in terms of end-user enjoyment, interest, and preferences, although diverging trends can be found when results are analysed according to excerpt and age group. Despite the translated version is assessed with higher median marks, the difference between the translated and the post-edited versions are rather low, and the results for the other analysed items are almost equal. According to the comprehension questionnaire, the translated version is better understood than the post-edited when both excerpts are taken into account. However, when the excerpts are analysed separately and age groups are taken into consideration, the results are contrasting. All in all, in none of the cases are the results statistically meaningful. The second experiment also took into account observational data from the dubbing session where the translations and post-edited texts were recorded. According to the data, translation is considered of better quality, although the differences are minimal in the case of the first excerpt.

On the whole, it can be claimed that the third secondary objective has been successfully achieved, as the quality of the post-edited and the translated texts have been assessed and compared. Furthermore, the results showed that the hypothesis upon which it relied was validated: the quality of the post-edited versions was significantly similar to the quality of the translations. However, the results of two experiments present no significant differences between the quality of translated and post-edited text, even though translation is mostly assessed more positively than post-editing. Furthermore, the achievement of this last secondary objective introduces another main contribution of this PhD: a quality assessment model to evaluate the quality of both translations and post-edited version of AVT field which combines a traditional quality assessment model with a user-reception study.

7.1.4. *Discussion of the Main Objective and Hypothesis*

The achievement of the three secondary objectives leads us to the main objective of this PhD: research whether MT might be successfully included into the process of translating documentaries of a certain subdomain –wildlife– through VO and OD AVT modalities.

The fulfilling of the first secondary objective of this PhD allowed us to determine the most important features to take into account if MT, and PE, were to be included into the process of translating wild life documentary films to be voiced-over and off-screen dubbing, while setting a framework for PE MT of documentaries to be voiced-over and off-screen dubbed. However, the solutions proposed to overcome the challenges were only partially used for the experiments, as it was considered an scenario were a domain specific MT engine could not be built, and hence, Google Translate was used to translate the excerpts that would be used throughout the experiments that form this PhD.

The second secondary objective was built upon the hypothesis that post-editing effort is lower than the effort of translating from scratch when translating wild life documentary films to be voiced-over and off-screen dubbed. While accomplishing the objective, the results of the carried out experiment proved the null-hypothesis of the study. The validation of the hypothesis allowed to go a step further in the fulfilling of the main hypothesis of this PhD, as it was proven that the level of effort decreases in when post-editing wild life documentaries.

However, in order to fully accomplish the main objective of this research, and thus validate its null-hypothesis, it was needed to prove that the quality of post-editing was significantly equal to the quality of translation. The third secondary objective of this PhD was fulfilled and the hypothesis on which it relied was validated: the quality of post-editing was significantly similar to the quality of translation. However, the quality of both translations and post-edited documents was rather low and translations were considered generally of better quality.

In conclusion, the accomplishment of the three secondary objectives and the validations of the two secondary hypothesis of this PhD lead us to claim that its main objective has been fulfilled. However, as the first experiment was carried out with MA students instead of professional translators, and the results of the quality assessment proved the quality of translation and post-editing was rather low, it is believed that the hypothesis of this PhD has only been proven null having into account the impact of the experimental conditions on the results of the studies carried out and further research is needed in order to determine if it would be actually possible in the industry. Furthermore, only one language pair has been used. Thus, further research is needed in order to fully accomplish the main objective of this PhD.

7.2. Main Contributions, Limitations and Further Research

Throughout the PhD and the conclusions it has been pointed out the main contributions, and the limitations of this research, which will be addressed in this section along with further research resulting from it.

This dissertation contributes to the different fields it relates to: AVT, MT and PE MT, and QA. All the contributions but one impact several of the fields at the same time and one of the contributions will have an impact on all the fields this PhD related to:

- Study on the inclusion of MT and PE MT and PE effort in VO:

Up until now research on VO, VO translation, and the translation of documentaries was devoted to reflect on its characteristics and particularities from both an academic (e.g. Espasa, 2006; Orero, 2006; Franco et al., 2010) and professional (e.g. Matamala, 2009) point of view. However, few applied studies had been done trying to use translation technologies into the process of translating documentaries, being this dissertation the first research that research the possibility of including MT and PE MT into the process of translating documentaries through VO.

In the fields of AVT and MT and PE MT, so far only research in subtitling (de Sousa et al, 2011; Etchegohyen et al., 2016) and audio description (Fernández-Torné et al., 2014), to a far lesser extent, included studies investigating the possibility of including MT and PE MT into the process. Hence, to the best of my knowledge, this PhD is the first research that intends to research the inclusion of MT into an oral AVT mode with non-controlled language.

Furthermore, as no research had been done investigating the inclusion of PE MT in VO or wild life documentary translation, this PhD represents the first incursion to the calculation of PE effort, and thus translation effort, for documentaries to be voiced-over.

All in all, this PhD contributes to both fields of study by expanding the framework researched so far.

- Division of effort depending on the task being performed:

Research on PE effort has focused on the three types of effort described by Krings (2001). However, even though studies have assessed effort by having into account if it is required to post-edit a whole text, a paragraph or sentences, the study presented in the second article is innovative, as it divides the effort depending on where the effort is put: in the main text while performing the task *per se* or in the audiovisuals or internet to perform a so called documentation task.

- AVT specifications in QA using MQM as framework:

In Translation Studies, QA has been broadly researched (e.g. House, 2007). In MT, both academy and industry are also preoccupied about this subject matter. However, in the field of AVT, QA has not been researched as broadly and reception studies are done in order to know whether the translations fulfill the expectations of the end-users.

The MQM framework has into account a set of specifications that contains a domain-specific type, which is to be customized by the user. As AVT and VO would be representatives of in-domain specifications, a set of specifications for VO, which could also be used for any other AVT oral modality, has been created for this dissertation. Such set of specifications is, thus, one of the main contributions of the PhD.

- Reception Study to assess quality:

Reception studies are used in AVT in order to determine if a certain translated product fulfills its purpose and is accepted by the end-user. In the field of AVT, almost no reception studies have been done involving VO and documentaries. It is innovative to use reception studies in MT or PE MT to assess quality. Therefore, this PhD contributes to the three fields of study by doing a reception study to assess the quality of post-edited wild life documentaries to be voiced-over and off-screen dubbed according to the end user.

This dissertation, however, also has some limitations. The first limitation of this research concerns the materials used throughout the studies. This research is only focused in one type of factual program: wildlife documentaries, and in a certain language combination: documentaries in English to be translated into Spanish. Furthermore, the chosen wildlife documentary film only contained formal register with some features of orality, as both the narrator and the experts used a high-level register but the speeches of the experts contained features of orality. The great amount of other factual programs that are usually translated by means of VO and OD opens the door to further research on the inclusion MT into the process of translation into other types of documentaries and factual programs that are translated by means of VO and OD. Additionally, more research needs to be done on the combination of registers within a same script, as many factual programs present different registers and an improvement on the performance of MT in documents with several registers might help increase the use of MT, decrease the post-editing effort, and better the levels of linguistic accessibility.

The second limitation of this research is related to the scenario chosen to build the experiments of this PhD. Despite the recommended solutions proposed in the first article of this PhD, which include the use of domain-specific engines and the inclusion of a pre-editing phase, this research has taken into account a scenario without the possibility of a customized domain specific MT engine and documents that needed no pre-editing phase. Therefore, another door opens for the MT researchers working on domain specific engines: as further research regarding the creation of domain specific MT engines for wildlife documentary films translations that could cope with not only specific terminology but also a combination of registers and, even, the automatic transcription of scripts would be needed.

The third limitation of this PhD concerns the particularities of the experiments. Although it was intended to simulate the current workflow, the first experiment was done in a laboratory environment with 12 MA students. The laboratory setup was selected over a more realistic current workflow environment for the sake of the well functioning of the experiment, as a keylogging software was needed in order to properly record the data to set the effort of post-editing and translation. Furthermore, MA students were used instead of professionals because of the lack of funding to involve professionals. However, the use of MA students had an impact on the results, as pointed out by the results on the quality assessment experiments. Hence, the experiments could be replicated with professional audiovisual translators in order to see whether the trend set by the

MA students is corroborated or if the results differ completely both in terms of productivity and quality. An experiment with professional audiovisual translators would also have some limitations, such as the expected lack of experience in post-editing compared to their wide experience as professional translators, which could a great impact on the results on the experiment on effort. Such limitation could be partially solved by doing a workshop on post-editing to the participants, as has been done in other studies such as Housley (2012). Nevertheless, making a workshop could lead to budget problems, as it could be hard to get funding for experiments with several participants that have to work in it for long hours, and also carry out a workshop before it. Furthermore, it would be difficult to find an expert on PE of AV products, as PE is still starting to be considered as an option in AVT industry and it is basically only applied to subtitling.

Currently, audiovisual translators who work on the translation of documentaries work in Word files. Therefore, in order to replicate the current workflow in a more real way, it was decided that the participants worked in Word files. Word files were also used because, to the best of my knowledge, no post-editing software that included the option of adding audiovisuals. Hence, there is a need to research on post-editing software that contains the possibility of including audiovisuals.

Even though there was online quality assessment software to evaluate translated, post-edited and machine translated texts, the texts were evaluated also in Word documents. It was decided to evaluate the translations and post-edited texts in Word documents because, to the best of my knowledge, the quality assessment software at the moment when the experiment was conducted did not allow to include new in-domain typologies into the set of specifications. Therefore, further research on the field of quality assessment regarding the inclusion of AVT typologies into the translation quality assessment needs to be carried out.

All in all, this section has presented the conclusions, as well as the main contributions, limitations, and further research that could arise from this PhD. At the beginning of this PhD I started pointing out the importance of linguistic accessibility and how MT can be helpful in order to make this type of accessibility increase. Thus, I want to finish this PhD with the same idea, while encouraging translators to approach MT without fear, and consider MT as a tool, rather than an enemy, that could help us achieve the ultimate goal of a translator: make a document linguistically accessible for everyone.

References

References

Allen, J. (2001). Postediting: an integrated part of a translation software program. *Language International*, *13*(2), 26-29.

— (2003). Post-editing. In H. Somers (ed). *Computers and Transaltion: A translator's guide* (35, p. 297-318). Amsterdam/Philadelphia: John Benjamins Publishing.

Almeida, G., & O'Brien, S. (2010). Analysing Post-Editing Performance: Correlations with Years of Translation Experience. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation* (no page numbers). St. Raphaël: EAMT. Retrieved from http://www.mt-archive.info /EAMT2010-Almeida.pdf [Consulted: 29 June 2016]

Armstrong, S., Caffrey, C., Flanagan, M., Kenny, D., & Way, A. (2006). Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation. In *MuTra-Multidimensional Translation Conference Proceedings. Audiovisual Translation Scenarios* (no page numbers). Copenhagen: MuTra. Retrieved from http://www.mt-archive.info/Aslib-2006-Armstrong.pdf [Consulted: 29 June 2016]

Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S., & Sadler, L. (1994). *Machine Translation: An Introductory Guide.* (p. 240). Oxford: NCC Blackwell.

Artegiani, I., & Kapsaskis, D. (2014). Template files: Asset or Anathema? A Qualitative Analysis of the Subtitles of *The Sopranos*. *Perspectives. Studies in Translatology*, *22*(3), 419-436. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0907676X.2013.833642 [Consulted: 13 July 2016]

Avramidis, E., Burchardt, A., Federmann, C., Popovic, M., Tscherwinka, C., & Torres, D. V. (2012). Involving Language Professionals in the Evaluation of Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 1127-1130). Istanbul: ELRA. Retrived from http://www.lrec-conf.org/proceedings/lrec2012/pdf/294_Paper.pdf [Consulted: 13 July 2016]

Avramidis, E., & Popovic, M. (2013). Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *Proceedings* *of the 8th Workshop on Statistical Machine Translation* (p. 329-336). Sofia: ACL. Retrived from http://www.aclweb.org/anthology/W13-22#page=349 [Consulted: 13 July 2016]

Aziz, W., de Sousa, S., & Specia, L. (2012). PET: a Tool for Post-Editing and Assessing Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 3982-3987). Istanbul: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf [Consulted: 29th June 2016]

Baldry, A., & Thibault, P.J. (2006). *Multimodal Transcription and Text Analysis. A Multimedia Toolkit and Coursebook* (p. 270). London/Oakville: Equinox.

Baker, M. (1992). *In Other Words. A Coursebook on Translation* (p. 330). London: Routledge.

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, *1*, 91-163.

Bouillon, P., & Spechbach, H. (2016). BabelDr: A Web Platform for Rapid Construction of Phrasebook-Style Medical Speech Translation Applications. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation* (4(2), p. 381). Riga: EAMT.

Bywood, L., Volk, M, Fishel, M., & Georgakopoulou, P. (2013). Parallel Subtitle Corpora and their Applications in Machine Translation and Translatology. *Perspectives: Studies in Translatology*. Special Issue: *Corpus Linguistics and Audiovisual Translation: in Search of a Integrated Approach*, *21*(4), 595-610. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0907676X.2013.831920 [Consulted: 13 July 2016]

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. (2011). The process of postediting: a pilot study. In *Proceedings of the 8th international NLPSC workshop* (p. 131-142). Frederiksberg: Samfundslitteratur. Copenhagen Studies in Language.

Carroll, J.B. (1966). An Experiment in Evaluating the Quality of Translation. *Mechanical Translation*, *9*(3-4), 55-66.

Carroll, M. (2004). Translation: A changing profession. *Translating Today*, *1*, 4-7.

Catford, J. (1965). *A Linguistic Theory of Translation* (31, p. 103). Oxford: Oxford University Press.

Ceausu, A., Tinsley, J., Zhang, J., & Way, A. (2011). Experiments on domain adaptation for patent machine translation in the PLuTO Project. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation* (no pages). Leuven: EAMT. Retrieved from http://doras.dcu.ie/16412/1/Experiments_on_Domain_Adaptation_for_Patent_Machine_Translation_in_the_PLuTO_project.pdf [Consulted: 13 July 2016]

Cettolo, M., Girardi, C., & Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation* (p. 261-268). Trento: EAMT. Retrieved from http://hnk.ffzg.hr/bibl/eamt2012/html/Papers/59.pdf [Consulted: 13 July 2016]

Chesterman, A. (2007). Bridge Concepts in Translation Studies. In M. Wolf, & A. Fukari (eds). *Translation Sociology. A new discipline under construction. Selected papers from the International conference on Translating and Interpreting as a Social Practice* (74, p. 171). Amsterdam/Philadephia: John Benjamins.

Chiaro, Delia. (2008). Issues of quality in screen translation. In D. Chiaro, C. Heiss, & C. Bucaria. *Between text and image: Updating research in screen translation* (p. 241-256). Amsterdam/Philadelphia: John Benjamins.

Dabbadie, M., Hartley, A., King, M., Miller, K.J., El Hadi, W.M., Popescu-Belis, A., Reeder, F., & Vanni, M. (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. In *Proceedings of the Workshop at the LREC 2002 Conference: Machine Translation Evaluation: Human Evaluators Meet Automated Metrics* (8, p. 8-16). Las Palmas: ELRA. Retrieved from https://www.semanticscholar.org/paper/Workshop-at-the-Lrec-2002-Conference-Machine-Palmas/7a4d645f83068c16c919289791c7fe7759677765/pdf [Consulted: 13 July 2016]

Daems, J., Macken, L., & Vandepitte, S. (2014). On the Origin of Errors: A Fine-Grained Analysis of MT and PE Errors and their Relationship. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'14)* (p. 62-

66). Reykjavik: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/ lrec2014/pdf/532_Paper.pdf [Consulted: 29 June 2016].

Darwish, A., & Orero, P. (2014). Rethorical dissonance of unsynchronized voices: Issues of voiceover in News Broadcasts. *Babel*, *60*(2), 129-144.

Daumé III, H., & Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers* (2, p. 407-412). Stroudsburg: ACL. Retrieved from http://delivery.acm.org/10.1145/2010000/2002819/p407-daume.pdf?ip=176.85.32.252&id=2002819&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=643378208&CFTOKEN=84140615&__acm__=1468445595_c1a6930d7f7323e154e356205d0d74cb [Consulted: 13 July 2016]

Díaz Cintas, J., & Orero, P. (2005). Screen Translation, Voice-over. Brown, K. (ed). *Encyclopedia of Languages* (p. 473-475). London: Elsevier.

Díaz Cintas, J., & Remael, A. (2007). *Audiovisual Translation: Subtitling* (p. 270). Manchester: St. Jerome.

Van Dijk, T.A. (1973). A Note on Linguistic Macro-Structures. P.T.G. Abraham, & P. Jordens (eds). *Linguistische Perspektiven* (p. 75-87). Tübingen: Niemeyer.

Van Dijk, T.A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Eckersley, H. (2002). Systems for Evaluating Translation Quality. *Multilingual Computing and Technology*, *3*(3), 39-42.

Englund Dimitrova, B. (2005). *Expertise and Explicitation in the Translation Process* (p. 418). Amsterdam/Philadelphia: John Benjamins.

Erickson, Megan. (2012). What's Lost (and Found) in Machine Translation. *Big Think*. Retrieved from: http://bigthink.com/think-tank/lost-and-found-in-machine-translation [Consulted: 9 July 2016]

Espasa, E. (2004). Myths about Documentary Translation. P. Orero (ed). *Topics in Audiovisual Translation* (p. 183-197). Amsterdam: John Benjamins.

Etchegoyhen, T., Bywood, L., Georgakopoulou, P., Fishel, M., Jiang, J., Loenhout, G., Pozo, A., Turner, A., Volk, M., & Maucec, M. (2014). Machine Translation for Subtitling: A Large-scale Evaluation. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)* (p. 46-53). Reykjavik: ELRA. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/463_Paper.pdf [Consulted: 13 July 2016]

European Comission. (2008). Multilingualism: an asset for Europe and a Shared commitment. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. {SEC(2008) 2443}{SEC(2008) 2444}{SeC(2008) 2445} /*COM(2008) 566 final*/

Fawcett, P. (1983). Translation modes and constraints. *The Incorporated Lingust*, *22*(4), 186-190.

Federmann, C., Melero, M., Pecina, P., & van Genabith, J. (2012). Towards Optimal Choice Selection for Improved Hybrid Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, *97*(1), 5–22. Retrieved from http://www.degruyter.com/view/j/pralin.2012.97.issue--1/v10108-012-0001-1/v10108-012-0001-1.xml [Consulted: 13 July 2016]

Fernández-Torné, A. (forthcoming). *Audio description and technologies. Study on the semi-automatisation of the transaltion and voicing of audio descriptions*. (PhD Dissertation). Universitat Autònoma de Barcelona.

Fernández-Torné, A., Matamala, A., & Ortiz-Boix, C. (2013) Enhancing Sensorial and Linguistic Accessiblity with Technology: Further Developments in the TECNACC and ALST projects. Paper presented at the *VI Media4All Conference* in Dubrovnik.

Fernández-Torné, A., & Matamala, A. (2015). Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan. *The Journal of Specialised Translation (JosTrans)*, *24*, 61-88.

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

Fields, P., Melby, A., Koby, G.S., Lommel, A., & Hague, D. (2014). What is quality? a management discipline and the translation industry get acquainted. *Tradumàtica: Tecnologies de la Traducció*, *12*, 404-412.

Fishel, M., Georgakopoulou, Y., Penkale, S., Petukhova, V., Rojc, M., Volk, M., & Way, A. (2012). From Subtitles to Parallel Corpora. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation* (p. 3-6). Trento: EAMT. Retrieved from http://hnk.ffzg.hr/bibl/eamt2012/EAMT-2012.pdf [Consulted: 29 June 2016].

Font Llitjós, A., Probst, K., & Carbonell, J. (2004). Error Analysis of Two Types of Grammar for the Purpose of Automatic Rule Refinement. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (p. 187-196). Washington, DC: AMTA. Retrieved from http://works.bepress.com/ jaim e ca rbonell/10/ [Consulted: 29 June 2016].

Font Llitjós, A., & Carbonell, J. (2004). The Translation Correction Tool: English-Spanish User Studies. In *Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC'14)* (no pages). Reykjavik: ELRA. Retrieved from http://repository.cmu.edu/cgi/viewcontent.cgi? article=1430and cont ext=isr [Consulted: 29 June 2016].

Franco, E.P.C. (2000). Documentary Film Translation: A Specific Practice? A. Chesterman, N. Gallardo San Salvador, & Y. Gambier (eds). *Translation in Context: Selected Contributions from the EST Congress* (p. 233-242). Amsterdam: John Benjamins.

—. (2001a). Inevitable Exoticism: The Translation of Culture-Specific Items in Documentaries. F. Chaume, & R. Agost (eds). *La traducción en los medios audiovisuales. Estudios sobre la Traducción*, (7, p. 177-181). Castelló de la Plana: Publicacions de la Universitat Jaume I.

—. (2001b). Voiced-over Television Documentaries: Terminological and Conceptual Issues for their Research, *Target*, *13*(2), 289-304.

Franco, E., Matamala, A., & Orero, P. (2010). *Voice-over Translation: an Overview* (p. 249). Bern: Peter Lang.

Freitag, M., Peitz, S., Wuebker, J., Ney, H., Durrani, N., Huck, M., Koehn, P., Ha, T.L., Niehues, J., Mediani, M., & Herrmann, T. (2013). EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *Proceedings of the 10th International Workshop for Spoken Language Translation (IWSLT2013)* (8, no pages). Heidelberg: KIT. Retrieved from http://workshop2013.iwslt.org/downloads/EUBRIDGE_MT_Text_Translation_of_Talks_in_the_EUBRIDGE_Project.pdf [Consulted: 29 June 2016]

Gambier, Y. (1996). La raduction audiovisuelle un genre nouveau? In Y. Gambier (ed). *Les tranferts linguistiques dans les médias audiovisuels* (p. 7-12). Villeneuve d'Ascq: Presses Universitaires du Septentrion.

— (ed). (1997). *Language transfer and audiovisual communication bibliography* (2nd ed.). Turku: Unipaps.

—. (1998). *Translating for the Media*. Turku: University of Turku.

—. (2008). Recent developments and challenges. D. Chiaro, C. Heiss, & C. Bucaria (eds). *Between text and image: Updating research in screen translation* (78, p. 11-34). Amsterdam/Philadelphia: John Benjamins.

Garcaraz, M. (2006). Poliskie tlumaczenia filowe. *The Journal of Specialised Translation (JosTrans)*, 5, 110-119.

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

García Luque, F. (2011). De cómo 'domesticar' un documental de divulgación científica en el proceso de traducción. Estudio de la versión en español de *L'Odyssée de l'espèce*. *SENDEBAR,* 22, 235-263.

Gaspari, F. (2001). Teaching machine translation to trainee translators: A survey of their knowledge and opinions. In *Proceedings of the MT Summit VIII Workshop on Teaching Machine Translation*, (no pages). Santiago de Compostela: EAMT.

Georgakopoulou, Y. (2010). Challenges for the Audiovisual Industry in the Digital Age: Accessibility and Multilingualism. *MetaNet Forum* (no pages). Brussels: META. Retrieved from http://www.meta-net.eu/events/meta-forum-2010/slides/META-FORUM2010_Georgakopoulou.pdf [Consulted: 29 June 2016]

Georgakopoulou, P., & Bywood, L. (2014). MT in subtitling and the rising profile of the post-editor. *Multilingual Computing, 25*(1), 24-28.

Gerlach, J., Porro Rodriguez, V., Bouillon, P., and Lehmann, S. (2013). Combining Pre-editing and Post-editing to Improve SMT of User-Generated Content. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice* (p. 45-53). Nice: EAMT. Retrieved from http://www.mt-archive.info/10/MTS-2013-W2-Gerlach.pdf [Consulted: 29 June 2016]

Giménez Linares, J.Á. (2008). *Empirical machine translation and its evaluation*. (PhD Dissertation). Universitat Politècnica de Catalunya.

Gadamer, H.G. (1960). *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik* (p. 495). Tübingen: J.C.B. Mohr.

Guerberof Arenas, A. (2009). Productivity and quality in MT post-editing. In *Proceedings of the MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT* (no pages). Ottawa: AMTA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.575.5398&rep=rep1&type=pdf [Consulted: 13 July 2016]

—. (2012). *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. (PhD Dissertation). Universitat Rovira i Virgili.

—. (2013). What do professional translators think about post-editing? *The Journal of Specialised Translation (JosTrans), 19*, .

Hague, D., Melby, A., & Zheng, W. (2011). Surveying translation quality assessment: A specification approach. *The Interpreter and translator trainer*, 5(2), 243-267. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13556509.2011.10798820 [Consulted: 13 July 2016]

Hansen, G. (2008). The speck in your brother's eye–the beam in your own. *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile, 80*, 255-277.

Hardmeier, C., & Volk, M. (2009). Using linguistic annotations in statistical machine translation of film subtitles. In *Proceedings of the 17th Nordic Conference of*

Computational Linguistics (NODALIDA 2009) (*16*, 57-64). Odense: Tartu University Library.

Hatim, B., & Mason, I. (1997). *The Translator as Communicator* (p. 215). London: Routledge.

Hurtado Albir, A. (2001). *Traducción y traductología* (p. 695). Madrid: Cátedra.

—. (2007). Compentence-Based Curriculum Design for Training Translators. *The Interpreter and Translator Trainer*, *1*(2), 163-195.

House, J. (2006). Text and context in translation. *Journal of Pragmatics*, *38*(3), 338-358.

Hutchins, W.J. (1986). *Machine Translation: Past, Present, Future* (p. 382). Chichester: Ellis Horwood.

Hutchins, J., & Somers, H. (1992). *An Introduction to Machine Translation* (p. 362). London: Academic Press.

Isabelle, P., Goutte, C., & Simard, M. (2007). Domain adaptation of MT systems through automatic post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation* (p. 255-261). Prague: ACL.

Joscelyne, A. (2009). *LSPs in the MT Loop: Current Practices, Future Requirements*. TAUS Report. Retrieved from https://www.taus.net/think-tank/reports/ translate-reports/lsps-in-the-mt-loop-current-practices-future-requirements [Consulted: 29 June 2016]

Kapsaskis, D. (2011). Professional Identity and Training of Translators in the Context of Globalisation: the Example of Subtitling. *The Journal of Specialised Translation (JosTrans)*, *16*, 162-184.

Kaufmann, F. (1995). Formation à la traduction et à l'interprétation pour les medias audiovisuels. In Y. Gambier (ed). *Communication audiovisuelle et transferts linguistiques. Audiovisual communication and language tranfers*. Special issue of *Translatio*, *14*(3-4), 431-442.

—. (2004). Un exemple d'effet pervers de l'uniformisation linguistique dans la traduction d'un documentarie: De l'hebreu des immigrants de "Saint'Jean" au français normatif d'ARTE. *Meta*, *49*(1), 148-160.

King, M., Popescu-Belis, A., & Hovy, E. (2003). FEMTI: Creating and Using a Framework for MT Evaluation. In *Proceedings of Machine Translation Summit IX* (p. 224-231). New Orleans: AMTA.

Kohen, P., & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation* (p. 102-212). New York City: ACL. Retrieved from http://homepages.inf.ed.ac.uk/pkoehn/publications/shared-task-wmt2006.pdf [Consulted: 29 June 2016]

Kohen, P., & Hoang, H. (2007). Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (p. 868-876). Prague: ACL. Retrieved from http://homepages.inf.ed.ac.uk/pkoehn/publications/emnlp2007-factored.pdf [Consulted: 29 June 2016]

Koller, W. (1987). Zum Gegenstand der Übersetzungstheorien. In R. Arntz, & G. Thome (eds). *Übersetzungswissenschaft. Ergebnisse und perspektiven* (p. 19-30). Tübingen: Narr.

Koponen, Maarit. (2015). Correctness of Machine Translation: a Machine Translation Post-editing Task. Paper presented at the *3rd MOLTO Project Meeting*. Retrieved from http://www.molto-project.eu/sites/default/files/molto_20110902_mkoponen.pdf [Consulted: 29 June 2016]

Koponen, M. (2016). *Machine Translation Post-editing and Effort. Empirical Studies on the Post-editing Process*. (PhD Dissertation). University of Helsinki.

Kovačič, I. (1995). Reception of subtitles. The non-existent ideal viewer. *Translatio*, *14*(3-4), 376-383.

—. (1998). Language in the media: A new challenge for translator training course. In. Y. Gambier (ed). *Translating for the media* (p. 123-129). Turku: Centre for Translation and Interpreting.

Krakovska, D. (2004). Simultaneous use of voice-over and subtitles for bilingual audiences. *Translating Today*, *1*, 25-27.

Krein-Kühle, M. (2014). Translation and Equivalence. In J. House. *Translation: A Multidisciplinary Approach* (p. 15-35). Palgrave: Macmillan.

Krings, H.P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*, 5. Kent: Kent State University Press.

Kupsch-Losereit, S. (1988). Die Übersetzung als soziale Praxis. Ihre Abhängigkeit vom Sinn- und Bedeutungshorizont des Rezipienten. *Fremdsprachen Lehren und Lernen*, *17*, 203-216.

Läubli, S., Fishel, M., Volk, M., & Weibel, M. (2013). Combining Statistical Machine Translation and Translation Memories with Domain Adaptation. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (85, p. 331-341). Oslo: Linköping University Electronic Press. Retrieved from http://stp.lingfil.uu.se/nodalida/2013/pdf/NODALIDA30.pdf [Consulted: 29 June 2016]

Lacruz, I., Denkowski, M., Lavie, A. (2014a). Cognitive Demand and Cognitive Effort in Post-Editing. In *Proceedings of the third Workshop on Post-Editing Technology and Practice* (p. 73-84). Vancouver: AMTA.

Lacruz, I., Denkowski, M., Lavie, A., & Dyer, C. (2014b). Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter. In *Proceedings of the Workshop on Human and Computer-asisted Translation* (p. 72-77). Gothengurg: ACL. Retrieved from http://anthology.aclweb.org/W/W14/W14-03.pdf#page=82 [Consulted: 13 July 2016]

Lavie, A., & Denkowski, M.J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105-115.

Leijten, M., & van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, *30*(3), 358–392.

León, B. (1998). Science Popularisation through Television Documentary: A Study of the Work of British Wildlife Filmmaker David Attenborough. In N. Sanitt (ed). *Proceedings of the 5th International Conference of Science and Technology* (p. 17-19). Berlin: The Pantaneo Forum.

Lison, P., & Tiedermann, J. (2016). Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th Language Resources and Evaluation*

Conference (LREC'16) (p. 923-929). Portoroz: ELRA. Retrieved from http://stp.lingfil.uu.se/~joerg/paper/opensubs2016.pdf [Consulted: 13 July 2016]

Lommel, A (2013). *Multidimensional Quality Metrics*. Paper presented at the *META-FORUM 2013*. Berlin: META. Retrieved from http://www.meta-net.eu/events/meta-forum-2013/talks/arlelommel.pdf [Consulted: 29 June 2016]

—. (ed). (2014) *Multidimensional Quality Metrics (MQM) Specifications*. Retriebed from <http://www.qt21.eu/mqm-definition/mqm-spec-2014-02-14.html> [Consulted: 29 June 2016]

Luyken, G.M., Herbst, T., Langham-Brown, J., Reid, H., & Spinhog, H. (1991). *Overcoming language barriers in television: Dubbing and subtitling for the European audience* [Media Monographs], 13. Manchester: The European Institute for the Media.

Mariana, V.R. (2014). *The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment*. (PhD Dissertation). Brigham Young University.

Martínez, L.G. (2003). *Human translation versus machine translation and full post-editing of raw machine translation output*. (MA Dissertation). Dublin City University. Retrieved from http://sceuromix.com/enlaces/MASTER%20IN%20TRANSLATION%20STUDIES%20BY%20LORENA%20GUERRA-2003.pdf [Consulted: 10 May 2016]

Matamala, A. (2002). *La traducción para voice-over: Online Module for the Master's Degree in Audiovisual Translation*. Barcelona: Universitat Autònoma de Barcelona.

—. (2004). Teaching Voice-over Translation. In *Languages and the Media: New Markets, New Tools. Conference Proceedings* (p. 24-26). Berlin: ICWE.

—. (2008). Teaching Voice-over Translation: A Practical Approach. In J. Díaz Cintas (ed). *The Didactics of Audiovisual Translation* (p. 231-262). Amsterdam: John Benjamins.

—. (2009a). Main Challenges in the Translation of Documentaries. In J. Díaz Cintas (ed). *New Trends in Audiovisual Translation* (p. 109-120). Bristol: Multilingual Matters.

—. (2009). Translating Documentaries: from Neanderthals to the *Supernanny*. *Perspectives: Studies in Translatology*, *17*(2), 93-107.

—. (2010). Terminological Challenges in the Translation of Science Documentaries: a Case-Study. *Across Languages and Cultures, 11*(2), 255-272.

—. (2015). The ALST project: technologies for audiovisual translation. In *Proceedings of the 37th Conference Translating and the Computer* (p. 79-89). London: AsLing.

Matamala, A., Fernández-Torné, A., & Ortiz-Boix, C. (2012). Technology and AD: The TECNACC Project. In *Languages and the Media 2012 Conference*. Berlin: ICWE. Retrieved from http://ddd.uab.cat/pub/presentacions/2012/117159/fernandez_matamala_ortiz_berlin2012.pdf [Consulted: 16 May 2016]

Matamala, A., & Ortiz-Boix, C. (Forthcoming). *Accessibility and Multilingalism: An Exploratory Study on the Machine Translation of Audio Descriptions*.

van der Meer, J., & Ruopp, A. (2014). MT Market Report 2014. *TAUS-Enabling Better Translation*.

Melby, A.K. (2006). MT+ TM+ QA: The future is ours. *Tradumàtica: traducció i tecnologies de la informació i la comunicació, 4,* no pages. Retrieved from http://www.raco.cat/index.php/Tradumatica/article/viewArticle/56004/0 [Consulted: 13 July 2016]

Melby, A.K., Fields, P., Koby, G.S., Lommel, A., & Hague, D. (2014). Defining the Landscape of Translation. *Tradumàtica: traducció i tecnologgies de la informació i la comunicació, 12*, 392-403. Retrieved from http://ddd.uab.cat/record/130154 [Consulted: 13 July 2016]

Melero, M., Oliver, A., & Badia, T. (2006). Automatic Multilingual Subtitling in the eTITLE Project. In *Proceedings of ASLIB Translating and the Computer* (28, no pages). London: ASLIB. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1 .1.107.6011andrep=rep1andtype=pdf [Consulted: 29 June 2016]

Moreau, H. (1998). L'interprétation sur la chaîne Arte. In Y. Gambier (ed). *Translating for the Media*, 225-229. Turku: Centre for Translation and Interpreting.

Nakov, P. (2008). Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Parapharsing, Tokenization and Recasing. In *Proceedings of the Third Workshop on Statistical Machine*

Translation (p. 147-150). Columbus: ACL. Retrieved from http://dl.acm.org/citation.cfm?id=1626414 [Consulted: 29 June 2016]

Nida, E.A. (1964). *Toward a Science of Translation: with special reference to principles and procedures involved in Bible translating* (p. 331). Leiden: Brill.

Nida, E.A. & Taber, C.R. (1969). *The Theory and practice of Translation* (p. 229). Leiden: Brill.

Nord, C. (1991). Text Analysis in Translator Training. In C. Dollerup, & A. Loddegaard (eds). *Teaching Translation and Interpreting: Training, Talent, and Experience* (p. 39-47). Amsterdam: John Benjamins.

—. (2014). *Translating as a Purposeful Activity: Functionalist Approaches Explained* (p. 154). Manchester: St Jerome Publishing.

O'Brien, S. (2004). Machine translatability and post-editing effort: How do they relate. In *Proceedings of the 26th International Conference on Translating and the Computer* (no pages). Edinburgh: ICSE. Retrieved from https://www.researchgate.net/profile/Sharon_Brien/publication/228784412_Machine_Translatability_and_Post-Editing_Effort_How_do_they_relate/links/54788bf10cf293e2da2b26d7.pdf [Consulted: 13 July 2016]

—. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation, 19*(1), 37-58.

—. (2006). Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. *Across Languages and Cultures, 7*(1), 1-21.

—. (2010). Introduction to Post-Editing: Who, What, How and Where to Next. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas* (no pages). Denver: AMTA. Retrieved from http://www.mtarchive.info/AMTA-2010-OBrien.pdf [Consulted: 29 June 2016]

—. (2011). Towards predicting post-editing productivity. *Machine Translation, 25*(3), 197-215.

Offersgaard, L., Povlsen, C., Almsten, L.K., & Maegaard, B. (2008). Domain specific MT in use. In *Proceedings of the 12th Annual Conference of the European Association for*

Machine Translation (p. 150-159). Hamburg: EAMT. Retrieved from http://curis.ku.dk/ws/files/9909035/EAMT-2008-Offersgaard.pdf [Consulted: 13 July 2016]

O'Hagan, M. (2007). Impact of DVD on Translation Language Options as an Essential Add-On Feature. *Convergence: The International Journal of Research into New Media Technologies*, *13*(2), 157-168.

Orero, P. (2004). The Pretended Easiness of Voice-over Translation of TV. *The Journal of Specialized Translation (JosTrans)*, *2*, 76-96.

—. (2006). Synchronization in Voice-Over. In Bravo, J.M. (ed). *Aspects of Translation* (p. 255-264). Valladolid: Universidad de Valladolid.

—. (2007). Voice-over: A Case of Hyper-Reality. In *EU High Level Scientific Conference Series. MUTRA Proceedings* (no pages). Copenhagen: ATRC. Retrieved from http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Orero_Pilar.pdf [Consulted: 29-June-2016].

Orero, P., & Matamala, A. (2007). Accessible opera: Overcoming linguistic and sensorial barriers. *Perspectives: Studies in Translatology*, *15*(4), 262-277.

Ortiz-Boix, Carla. (2012). *Tecnologies per a l'audiodescripció: estudi sobre l'aplicació de la traducció automàtica i la síntesi de parla a l'audiodescripció en castellà*. (MA Dissertation). Universitat Autònoma de Barcelona.

—. (Forthcoming). Post-Editing Wildlife Documentaries: Challenges and Possible Solutions. *Hermeneus, 18*.

Ortiz-Boix, C., & Matamala, A. (forthcoming). Post-Editing Wildlife Documentary Films: a new possible scenario? *Journal of Specialized Translation (JosTrans), 26*, 187-210.

Ortiz-Boix, C., & Matamala, A. (forthcoming). Assessing the Quality of Post-Edited Wildlife Documentaries. *Perspectives. Studies in Translatology*.

Ortiz-Boix, C., & Matamala, A. (2015). Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: a Three-Level Approach. In O'Brien, S. & Simard, M. (eds). *Proceedings of the 4th Workshop on Post-Editing Technology and Practice (WPTP4)* (p. 16-30). Miami: AMTA. Retrieved from http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf [Consulted: 13 July 2016]

Ortiz-Martínez, D., Sanchís, G., Casacuberta, F., Alabau, V., Vidal, E., Benedí, J.M., González-Rubio, J., Sanchís, A., & González, J. (2012). The CASMACAT Project: The Next Generation Translator's Workbench. In *Proceedings of the 7th* Jornadas en tecnologia del Habla *and the 3rd Iberian SLTech Workshop (IberSPEECH)* (p. 326-334). Madrid: Universidad Autónoma de Madrid. Retrieved from http://www.casmacat.eu/uploads/Main__/iberspeech1.pdf [Consulted: 29 June 2016]

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (p. 311-318). Stroudsburg: ACL. Retreived from http://delivery.acm.org/10.1145/1080000/1073135/p311-papineni.pdf?ip=176.85.32.252&id=1073135&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=643378208&CFTOKEN=84140615&__acm__=1468451020_756ba83ab9f75e6b1d06a593338750ed [Consulted: 13 July 2016]

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

Pönniö, K. (1995). Voice over, narration et commentaire. Gambier, Y. (ed). *Communication audiovisuelle et transferts linguistiques/Audiovisual Communication and Language Transfer*. Special issue of *Translatio*, 303-307. Strasbourg: Fédération Internationele des Traducteurs (FIT).

Popović, M., & Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, *37*(4), 657-688.

Pozo, A. del, Bywood, L., Georgakopoulou, P., Etchegoyhen, T., Fishel, M., Jiang, J., Loenhout, G., Spiliotopoulos, D., Turner, A., & Maucec, M. (2012). *SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report*. Retreived from http://cordis.europa.eu/fp7/ict/language-technologies/docs/sumat-annual-report -2012.pdf [Consulted: 29 June 2016]

Pym, P. J. (1990). Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system. *Translating and the Computer*, *10*, 80-96.

Ray, R. (ed). (2004). *LISA Best Practice Guides. Implementing Machine Translation*. Retrieved from http://www.translationoptimization.com/papers/DillingerLommel_MT_BPG.pdf [Consulted: 29 June 2016]

Reiss, K. (1978). Übersetzungstheorie und Praxis der Übersetzungskritik. In F. Königs (ed). *Übersetzungswissenschaft und Fremdsprachenunterricht* (p. 71-93). München: Goethe-Institut.

Reiss, K., and Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie* (p. 254). Tübingen: Niewmeyer.

Remael, Alice. (2010). Audiovisual Translation. In Y. Gambier, & L. van Doorsalaer *Handbook of Translation Studies* (p. 12-17). Amsterdam: John Benjamins.

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

Robinson, D. (2004). *Becoming a translator: An introduction to the theory and practice of translation* (p. 235). London: Routledge.

Roturier, J. (2006). *An Investigation into the impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical.* (PhD Disseratation). Dublin City University.

Sánchez, D. (2004). Subtitling Methods and Team-Translation. In P. Orero (ed). *Topics in Audiovisual Translation* (56, p. 9-17). Amsterdam/Philadelphia: John Benjamins.

Savoy, T. (1968). *The Art of Translation* (p. 322). Boston: The Writer.

Scannel, P. (1996). *Radio, television and modern life* (p. 204). Oxford: Blackwell.

Sepielak, K. (2014). Translation techniques in voiced-over multilingual feature movies. *Lingüística Antverpiensia, New Series – Themes in Translation Studies*, *14*, 251-272.

Sepielak, K., & Matamala, A. (2014). Synchrony in the voice-over of Polish fiction genres. *Babel*, *60*(2), 145-163.

Sertan, V., Bouillon, P., & Gerlach, J. (2014). A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation. In *Proceedings of* *the 8th International Conference on Language Resources and Evaluation (LREC'14)* (p. 1793-1799). Reykjavik: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/676_Paper.pdf [Consulted: 29 June 2016]

Shreve, G.M., Lacruz, I., & Angelone, E. (2011). Sight Translation and Speech Disfluency: Performance Analysis as Window to Cognitive Translation Processes. In c. Alvstad, A. Hild, & E. Tiselius (eds). *Methods and Strategies of Process Research* (p. 121-146). Amsterdam: John Benjamins.

Snover, M., Dorr, B.J., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas* (p. 223-231). Cambridge: AMTA.

Snover, M., Madnani, N., Dorr, B.J., & Schwartz, R. (2009). Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (p. 259-268). Athens: ACL.

Sousa, S., Aziz, W., & Specia, L. (2011). Assessing the Post-editing Effort for Automatic and Semi-automatic Translations of DVD subtitles. In *Proceedings of Recent Advances in Natural Language Processing Conference* (p. 97-103). Hissar: ACL. Retrieved from http://aclweb.org/anthology/R11-1014 [Consulted: 29 June 2016]

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation* (p. 73–80). Leuven: EAMT.

Steiner, G. (1975). *After Babel: Aspects of Language and Translation* (p. 538). Oxford: Oxford University Press.

Sutter, N., & Depraetere, I. (2012). Post-edited translation quality, edit distance and fluency scores: report on a case study. Paper presented at the *Journée d'études. Traduction et qualité Méthodologies en matière d'Assurance Qualité* Conference. Lille: Université Lille. Sciences humaines et sociales.

Tatsumi, M., & Roturier, J. (2010). Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship. In *Proceedings of the*

Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC10) (p. 43-51). Denver: AMTA.

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

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. In *Proceedings of the European Conference on Speech Communication and Technology* (p. 2667–2670). Rhodes: ISCA.

Toury, G. (1985). *In Search of a Theory of Translation*. Tel Aviv: Tel Aviv University.

Uszkoreit, H., & Lommel, A. (2013). *Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment*. Retrieved from http://www.qt21.eu/launchpad /sites/default/files/MQM.pdf [Consulted: 29 June 2016]

Valor Miró, J.D., Silvestre-Cerdà, J.A., Civera, J., Turró, C., & Juan, A. (2015). Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication, 74*, 65-75.

Veale, T., & Way, A. (1997). Gaijin: A bootstrapping, template-driven approach to example-based MT. In *Proceedings of the New Methods in Natural Language Processing (eMNLP97)* (p. 239-244). Providence: ACL.

Venuti, L. (1995). *The Translator's Invisibility. A History of Translation* (p. 319). London: Routledge.

Vilar, D., Xu, J., d'Haro, L.F., & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)* (p. 697-702). Genoa: ELRA. Retrieved from http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf [Consulted: 29 June 2016]

Volk, M. (2008). The Automatic Translation of Film Subtitles. A Machine Translation Success Story? *Journal for Language Technology and Computational Linguistics, 23*(2), 113-125.

Volk, M., Sennrich, R., Hardmeier, C., & Tidström, F. (2010). Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+ICNGL Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC'10)* (p. 53-62). Luxemburg: JEC. Retrieved from http://www.mt-archive.info/10/JEC-2010-Volk.pdf [Consulted: 29 June 2016]

Waibel, A. (2012). *EU-Bridge Newsletter*. (1st Edition, p. 8). Retrieved from http://project.eu-bridge.eu/img/Newsletter_1_edition.pdf [Consulted: 29 June 2016]

Williams, M. (2001). The Application of Argumentation Theory to Translation Quality Assessment. *META, 46*(2), 327-344.

Zhechev, V. (2014). Analysing the Post-Editing of Machine Translation at Autodesk. In *Post-Editing of Machine Translation: Processes and Applications* (p. 2-23). Newcastle upon Tyne: Cambridge Scholars Publishing.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)* (p. 3530-3534). Portoroz: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/pdf/1195_Paper.pdf [Consulted: 13 July 2016]

Zinik, Z. (2006). Freelance. *Times Literary Supplement*, *14*, 53-60.

Filmography

Filmography

National Geographic (ed). (2009). Must Watch: A lioness adopts a baby antelope. *Unlikely Animal Friends*. Episode: Odd Couples.

Fothergill, A. (2006). *Planet Earth*. Producer: BBC Natural History, UK.

Staiton, J. (1997-2004). *The Crocodile Hunter*. Producer: The Best Picture Show Company, Australia.

Annexes

Annex A

A1. 1st Article

MACHINE TRANSLATION AND POST-EDITING IN WILDLIFE DOCUMENTARIES: CHALLENGES AND POSSIBLE SOLUTIONS

Traducción automática y posedición para documentales de naturaleza: desafíos y posibles soluciones

Carla ORTIZ-BOIX[1]
Universitat Autònoma de Barcelona

ABSTRACT: This article presents some of the challenges that may have to be overcome in order to introduce Machine Translation (MT) into the process of translating wildlife documentary films. Until now, MT has mainly been applied to written general and specialized texts. However, in the past few years, EU-financed projects have started to work in the field of audiovisual translation with the aim to introduce MT into subtitling. It has already been proven that post-edited machine translated subtitles can reach the appropriate quality levels. Nevertheless, in the case of documentaries, not only subtitling but also voice-over and off-screen dubbing can be found in countries where subtitling is not the main audiovisual transfer mode. Therefore, similar research in voice-over and off-screen dubbing is believed to be worthy. This article aims to describe the challenges of machine translating documentary scripts by presenting a preliminary analysis on the translations produced by MT engines. Firstly, an overview of the characteristics of voice-over and off-screen dubbing is provided, as well as a brief review dealing with MT and post-editing in audiovisual translation. Next, the methodology used to carry out the analysis of both a corpus of documentary scripts and a corpus of machine translations of documentary scripts is explained. Finally, before summarizing new potential avenues of research, the challenges that may have to be faced in order to achieve high quality translations of documentary scripts using MT are pointed out, the results of the analysis are presented, and some possible solutions are suggested.

Key words: post-editing, audiovisual translation, voice-over, off-screen dubbing, documentaries, machine translation, pre-editing.

RESUMEN: Este artículo presenta algunos de los desafíos que pueden presentarse si introducimos traducción automática (TA) en el proceso de traducción de documentales de naturaleza. Hasta ahora, TA se ha usado para traducir textos escritos de carácter general y especializado. A pesar de ello, en los últimos años, proyectos financiados por la UE han empezado a trabajar en el ámbito de la traducción audiovisual con el objetivo de usar TA para traducir subtítulos y ya se ha demostrado que los subtítulos poseditados pueden llegar a niveles de calidad adecuados. Pero los documentales no solo se pueden traducirse mediante subtítulos ya que, en países donde la subtitulación no es el principal modo de transferencia audiovisual, se usan voces

[1] This article is part of the ALST Project (reference FFI-2012-31024), directed by Anna Matamala and funded by the Spanish Ministerio de Economía y Competititvidad, and also of research group TransMedia Catalonia (2014SGR27); and of the research carried out by Carla Ortiz-Boix under the supervision of Dr Anna Matamala within the PhD in Translation and Intercultural Studies (Department of Translation, Interpreting, and East Asian Studies) at Universitat Autònoma de Barcelona. It is also partially funded by the FI-DGR scholarship awarded by AGAUR (FI-DGR2013). The data was partially gathered and analysed during a research stay at Pangeanic translation company. I would also like to thank Manuel Herranz for welcoming me on a research stay at Pangeanic.

superpuestas y doblaje en *off* para hacerlo. Es por este motivo, pues, que creemos necesario investigar la introducción de TA para traducir documentales de naturaleza mediante voces superpuestas y doblaje en *off*. Este artículo describe los desafíos que conlleva traducir automáticamente guiones de documentales presentado un análisis preliminar de las traducciones producidas por distintos motores de traducción automática. En primer lugar aportamos una visión general de las características de las voces superpuestas y el doblaje en *off*, así como un breve resumen de anteriores investigaciones en las que se intenta introducir TA en el ámbito de la traducción audiovisual. A continuación presentamos la metodología usada para llevar a cabo el análisis de un corpus de guiones de documentales, por un lado, y de un corpus de traducciones automáticas de estos mismos guiones, por el otro. Finalmente, antes de resumir posibles nuevas investigaciones derivadas de este artículo, esclarecemos los posibles desafíos con los que podríamos encontrarnos para conseguir traducciones de guiones de documentales de calidad usando TA, presentamos los resultados de los análisis y sugerimos posibles soluciones a estos desafíos.

Palabras clave: posedición, traducción audiovisual, voces superpuestas, doblaje en *off*, traducción automática, preedición.

1. INTRODUCTION

Research on Machine Translation (MT) and post-editing (PE) has attracted great interest over the last decade, not only among Translation Studies scholars, but also among translation industry stakeholders. TAUS (Joscelyne 2009) market study indicates that 92.23% of the language server providers included in its study already use or intend to use MT and PE as part of their translation process. However, in the audiovisual translation (AVT) market, professional experiences in MT and PE are limited (Volk *et al* 2010) and industry voices in favour of MT are just beginning to be heard (Georgakopoulou 2010). Interest in academia has increased in recent years, focussing on the implementation of MT and PE in subtitling, in part due to EU-financed projects such as eTITLE (Melero *et al.* 2006) EU-Bridge (Waibel 2012) or SUMAT (Del Pozo *et al.* 2012). The promising results of these studies (Fishel 2012; Bywood *et al.* 2013; Freitag *et al.* 2013) have encouraged other researchers to study the inclusion of MT in other AVT modes such as audio description (Ortiz-Boix 2012; Fernández *et al.* 2013).

Inspired by existing research, I have started an investigation based on the hypothesis that MT can be successfully implemented when translating wildlife documentaries for oral transfer modes such as voice-over (VO) and off-screen dubbing. This research will assess the quality of MT output, and most importantly, the PE effort as compared to a standard human translation. However, before carrying out this experimental part of the research, I have considered it relevant to do a bibliographical survey and carry out a qualitative analysis on a corpus of documentaries, in order to point out the specific problems that will probably have to be addressed. As documentary films can deal with a wide variety of subjects, such as arts, health, history, music or wildlife, to mention but a few, and each topic has its own terminological specificities, a specific domain has been selected to narrow down the analysis: wildlife. This is due to the fact that there is a wide variety of wildlife documentary films; while some present species or ecosystems through beautiful images, the voice of a narrator, and sometimes, of experts (*Planet Earth* 2006), others are almost reality programs (*The Crocodile*

Hunter, 1997-2004). Furthermore, wildlife documentaries are frequent in TV's daily schedule – both in Spanish and English-speaking countries, illustratively on channels such as Animal Planet, BBC One, BBC Four, National Geographic Wild, La 2, and Canal Plus. The article aims to present the potential challenges arising from the use of MT engines and PE software when translating for VO and off-screen dubbing, two audiovisual transfer modes which can be often found in wildlife documentary films.

The article focuses on eight challenges and their possible solutions: (1) spotting, (2) synchronization, (3) access to audiovisual content, (4) variety on the script format, (5) register variety within a same script, (6) terminology, (7) errors and inaccuracies in the original script, and (8) linguistic inconsistencies in the original script. In order to identify the challenges, two approaches have been taken: on the one hand a bibliographical survey of existing literature on VO and off-screen dubbing has been conducted, and on the other, an analysis of two corpora, namely a corpus of wildlife documentary scripts in English and a corpus in Spanish. An error analysis of a corpus of 50 sentences machine translated using eight free online engines provides additional insight into the most common errors produced by MT engines.

The article is divided as follows: a short overview on the two transfer modes under analysis (VO and off-screen dubbing), as applied to the translation of documentaries, as well as a short review of previous MT and PE research within AVT are presented in sections 2 and 3 respectively. In section 4, the methodology used to identify the challenges is explained. Sections 5, 6 and 7 present the challenges: section 5 focuses on the challenges found in previous academic works, section 6 describes those derived from the analysis of corpora 1 and 2, and section 7 lists the challenges found through both the automatic and human evaluations of the corpus of 50 sentences. In section 8, possible solutions are proposed, and in the last section, conclusions and further research are presented.

2. VOICE-OVER AND OFF-SCREEN DUBBING

The branch of Translation Studies that deals with documentary films is AVT, which can be described as the field of Translation Studies concerned with the transfer of multimodal and multimedia texts into another language and/or culture (Baldry & Thibault 2006). Although there are many AVT transfer modes (subtitling, dubbing, audio description, surtitling, voice-over, subtitling for the deaf and hard of hearing, live subtitling, video-game localization, etc. [Remael 2010]) and almost all of them could be used in a documentary film, this article focuses only on off-screen dubbing and VO of wildlife documentary films, from English into Spanish. These two modes have been selected as they are the most used in open and closed TV channels in Spain, for instance, where it is common to find documentaries in which the narrator is re-voiced by using off-screen dubbing, whilst interviewees are rendered via VO. Although research in these transfer modes and genres initially received little attention, the trend has changed in recent years with some more works being published: Espasa (2004), Franco (2000, 2001a, 2001b), García Luque (2011), Matamala (2002, 2004, 2008, 2009a, 2009b), and Orero (2004, 2007).

Díaz Cintas and Orero (2006: 473) define voice-over as follows:

> Technique in which a voice offering a translation in a given target language is heard simultaneously on top of the SL voice. As far as the soundtrack of the original program is concerned, the volume is reduced to a low level that can still be heard in the background when the translation is being read. It is common practice to allow the viewer to hear the original speech in the foreign language at the onset of the speech and to reduce subsequently the volume of the original so that the translated speech can be inserted. The translation usually finishes several seconds before the foreign language speech does, the sound of the original is raised again to a normal volume and the viewer can hear once more the original speech.

According to Franco *et al.* (2010: 25), voice-over translation in factual programmes is said to help reproduce the feeling of reality, truth and authenticity that the original audiovisual product gives, which is supported both by visual evidence (images of events, people, documents and archival footage) and by verbal evidence (interviews with experts and witnesses). The delivery of VO does not usually show regional accents in the target text and does not generally reproduce specific oral features such as fluffs, hesitations or grammatical mistakes. Orero (2006) highlights the importance of three types of synchrony in VO: kinetic synchrony –the voice delivering the translation matches the body movements which can be seen on screen–, action synchrony –the voice delivering the translation matches the actions taking place on screen–, and voice-over isochrony –the translated message fits between the beginning and the end of the original speech, leaving some time before it starts and after it ends during which the original soundtrack is heard.

Off-screen dubbing, also termed commentary and narration by authors like Pönniö (1995), shares kinetic and action synchrony with voice-over but not voice-over isochrony. This is because the original voice is not heard but instead substituted by the target language. Additionally, VO is generally used for semi-spontaneous or spontaneous interviewees, whilst off-screen dubbing is usually applied to narrators with a planned discourse, and this also has implications in the language register.

Other shared features pointed out in the literature (Franco *et al.* 2010) are the lack of postproduction scripts or, if available, the poor quality of the transcriptions provided to the translators, which may contain linguistic errors and inaccuracies, etc. (see sections 6.2.4 and 6.2.5). Furthermore, as Matamala (2010) states, wildlife and scientific documentaries –the specific focus of this research– make use of a vast array of terminology, which might be a challenge for their translation (see section 6.2.3).

3. MACHINE TRANSLATION IN AUDIOVISUAL TRANSLATION

So far, implementing MT into the translation process has proven successful in limited domains, such as meteorology or finances, and when working with general texts, in which case MT is used for gisting purposes and for interpersonal communication (Ray 2004: 8-9). MT engines are becoming more and more domain-specific, which guarantees a better quality translation for the post-editors to work with (Läubli *et al.* 2013: 2). In the case of AVT, the implementation of MT is falling behind, as it has only been researched in subtitling (Melero 2006; Armstrong *et al.* 2006; Volk 2008; Bywood 2013), and to a far lesser extent, audio description (Ortiz-Boix 2012; Fernández *et al.* 2013).

Different approaches have been adopted to implement MT in the field of subtitling. Armstrong *et al.* (2006) have researched quality improvement when translating subtitles in the language pair English <> German with an EBMT engine with homogeneous data in comparison with an EBMT with heterogeneous data. The completed eTITLE Project (Melero 2006) intended to increase the efficiency of subtitling by automating various processes within its workflow, achieving a good BLEU score (36.9) in the English-Spanish combination.

Research in this field has also been carried out by Volk (2008), who has investigated whether it is feasible to use MT in subtitling by focusing on the language combination Danish <> English and checking three criteria: number of users, customer satisfaction, and long-term usage of the MT system. He concludes that it is feasible as the statistical MT based system reached high BLEU scores (average 57.3) and saved time in the translation process. Furthermore, he points out the possibility of adding pre-editing to control the language of the source documents so that the MT system is more competitive.

In the case of audio description, Ortiz-Boix (2012) presents a preliminary study on the application of MT to audio description process in the Catalan <> Spanish language pair. Although it is a preliminary study within the context of an MA dissertation, the first results are reassuring as the lowest BLEU score was 67.00. MT is envisaged by this researcher as a tool to increase accessibility in multilingual environments by working with closely related languages (Matamala *et al.* forthcoming).

Finally, the most recent project on the topic, SUMAT (Online Service for Subtitling by MT, see http://www.sumat-project.eu/), works with 14 different language pairs and initial BLEU results of 25.5 are promising (Bywood 2013). The project aims to provide not only automatic measures but also to test the human PE effort, an approach taken in general translation (De Almeida *et al.* 2010) but almost absent in AVT (Sousa *et al*. 2011).

To sum up, the existing results regarding the application of MT and PE to subtitling and audio description processes have compelled us to put forward the hypothesis that MT with PE could also be successfully implemented into the translation of documentary films. Before carrying out experimental research to prove this hypothesis, a qualitative analysis has been done to foresee possible challenges, as described in the next section.

4. METHODOLOGICAL CONSIDERATIONS

Two methodological approaches have been adopted: on the one hand, a bibliographical review, which has led us to identify three challenges (discussed in section 5), and on the other, an analysis of three corpora which has allowed to confirm some of the issues found in the bibliographical survey, and to add some new ones (see sections 6 and 7). The main features of the corpora and how they have been analysed are explained next.

4.1. CORPUS CREATION

In order to find the characteristics of documentary scripts that can impact MT and PE processes, 108 documentary scripts in English (original texts) and 92 in Spanish (translations) have been collected and analysed. Some of the documentaries (66) only contain a narrator to be revoiced using off-screen dubbing, whilst others (54) contain a narrator plus interviewees and spontaneous speech to be voiced-over. These scripts were divided into three corpora:

1. En-Doc: 108 English documentary scripts in English, containing 504,368 words in 13,426 sentences (see table 1).

2. Spa-Doc: 92 documentary scripts in Spanish containing 440,651 words in 7,053 sentences. 80 of them are human translations of the documentaries included in En-Doc, whilst the remaining 12 are also human translations whose original script is not included in the previous corpus (see table 1).

CORPUS	SCRIPTS	SEGMENTS	WORDS
En-Doc	108	13,426	504,368
Spa-Doc	92	7,053	440,651

Table 1. En-Doc & Spa-Doc Corpora

3. Bil-Doc: constituted by a random selection of 50 original English segments (meaning group of words, *i.e.* whole sentences or syntagmas the MT engine is fed with) next to their human translation and eight MTs into Spanish. It contains 6,592 words (633 English and 5,959 Spanish words), as shown in table 2:

CORPUS	LANGUAGE	SEGMENTS	WORDS
Bil-Doc	English	50	633
	Spanish	450	5,959

Table 2. En-Doc & Spa-Doc Corpora

The 50 random segments in English and their translations in Spanish were extracted from the 80 documentary scripts the Spanish translation of which was already available. Only text that has to be voiced –and therefore needs to be translated– was considered and additional information on the visuals or music –generally omitted from the translation but sometimes included in the scripts– was disregarded in this selection. The segments were translated using the English into Spanish free online MT engines that were found on the web, (of which there are only eight) when the analysis took place (table 3):

MT Engine	Website
Apertium	www.apertium.org/#translation
Bing	www.bing.com/translator
Google Translate	http://translate.google.com/
Lucy MT	www.lucysoftware.com/english/machine-translation/lucy-lt-kwik-translator-/
Promt	www.online-translator.com/
Reverso	www.reverso.net/text_translation.aspx?lang=ES
Systran	www.reverso.net/text_translation.aspx?lang=ES
Yandex	https://translate.yandex.com/

Table 3. MT Engines

4.2. CORPUS ANALYSIS

En-Doc and Spa-Doc corpora helped determine some of the challenges regarding wildlife documentary scripts' features. Both micro- and macro-structures of documentaries in English and Spanish were analysed and compared. Macro-structures are "the overall structures of a text" (Van Dijk 1973: 73), whilst micro-structures are understood as the connections between words and sentences within a text which become the basis for its general meaning (Van Dijk *et al*. 1983: 73).

1. *Macro-structure analysis in both En-Doc and Spa-Doc corpora*: a manual analysis of the script layout was carried out and divergences were found in the formatting of time codes and the inclusion of additional contents (description of visual information, details about the music heard, etc.). The results of this corpus-based bottom-up analysis, which was not based on any previous categorisation, were compared with the script layouts found in Franco *et al.* (2010). This analysis, the results of which can be found in sections 5.1 and 6.1, was carried out for both En-Doc and Spa-Doc corpus independently, and the results were then compared.

2. *Micro-structure analysis in both En-Doc and Spa-Doc corpora*: this analysis adopted a different approach, resting on a pre-established categorisation from previous literature. A list of categories (namely terminology, register, linguistic inconsistencies, inaccuracies and errors in the original script) was searched manually in the corpus in order to confirm or reject their presence, hence offering qualitative data through a top-down corpus-based analysis. This analysis, the results of which can be found in sections 5.2, 6.2, 6.3, 6.4, and 6.5., was carried out for both En-Doc and Spa-Doc corpora independently, and then results were compared.

3. *Analysis of the Bil-Doc corpus:* this corpus was used to confirm some of the previously found challenges regarding micro-structure, as well as to run a preliminary test on the possible application of MT to wildlife documentary films, and to determine the most common errors when machine translating wildlife documentary films. Therefore, an automatic and a human subjective evaluation were made.

In order to analyse the Bil-Doc corpus and to evaluate the translations, several steps were followed:

1. An automatic evaluation, the results of which can be found in section 7.1, was made using Asia Online software (www.asiaonline.net), providing BLEU and TER automatic measures of the eight MT engines' translations against the existing human translations.

2. A subjective assessment of the output from all eight MT engines was made by one researcher (results can be found in Section 7.2). All errors were marked and classified according to a table based on the Multidimensional Quality Metrics Error Typology (MQM) proposed by Uszkoreit *et al.* (2013). Quality assessment of human translations has been researched by many authors in translation studies (*e.g.* Hurtado Albir 2001; Williams 2001; Eckersley 2002; Hurtado Albir 2007; Nord 2014), who have proposed different categorizations of errors. However, they do not take into account the specificities of MT. This is why a categorization of errors specifically for MT output was considered the most appropriate for the presented assessment, as the analysed output was machine translated. Among all error categorizations available that asses MT output (*e.g.* Font Llitjós *et al.* 2004a; Koponen 2010), MQM was selected as a starting point because it is the most exhaustive and allows researchers to introduce domain-specific categories or erase unneeded categories. In any case, only categories regarding accuracy, issue, type and mechanical issues included in fluency were used for the purposes of this article as they are considered the most relevant (Uszkoreit *et al.* 2013). Table 4 lists all error categories used in this article:

ACCURACY	Terminology		A term is translated with a term other than the one expected for the domain or otherwise specified.
	Mistranslation		The target content does not accurately represent the source content.
		Overly Literal	The translation is overly literal
		False Friend	The translation has incorrectly used a word that is superficially similar to the source word.
		Sould not have been translated	Text was translated that should have been left untranslated.
		Date/time	Dates or times do not match between source and target.
		Unit conversion	The target text has not converted numeric values as needed to adjust for different units.
		Number	Numbers are inconsistent between source and target.
		Entity	Names, places or other "named entities" do not match.
	Omission		Content is missing from the translation that is present in the source.
	Addition		The target text includes text not present in the source.
	Untranslated		Content that should have been translated has been left untranslated.
FLUENCY	Spelling		Issues related to spelling of words.
		Capitalization	Issues related to capitalization.
		Diacritics	Issues related to the use of diacritics
	Typography		Issues related to the mechanical presentation of text. The category should be used for any typographical errors other than spelling.
		Punctuation	Punctuation is used incorrectly for the locale or style.
		Unpaired quote marks or brackets	One of a pair of quotes or brackets is missing from the text.
	Grammar		Issues related to the grammar or syntax of the text, other than spelling and orthography.

		Morphology	There is a problema in the internal construction of a word.
		Part of speech	A word is the wrong part of speech.
		Agreement	Two or more words do not agree with respect to case, number, person or other grammatial features.
		Word order	The word order is incorrect.
		Function words	A function word is used incorrectly.
	Unintelligible	The exact nature of the error cannot be determined. Indicates a major break down in fluency.	

Table 4. Used Metrics for human evaluation based on MQM (Uszkoreit *et al*. 2013)

After categorising the errors by marking and processing them with an Excel spreadsheet, the results of each MT engine were analysed and compared.

Before introducing the results of the analyses, namely the foreseen challenges if MT is included in the process of translating wildlife documentaries to be voiced-over and off-screen dubbed, a summary of the methodology –including in which Section the results can be found– is presented in table 5.

Approach	CORPUS	Type of Analysis	Results in…
(a) Bibliographical review			Section 5
(b) Corpus analysis	EN-DOC corpus	Micro- and macro-structure analysis	Section 6
	SPA-DOC corpus	Micro- and macro-structure analysis	Section 6
	BIL-DOC corpus	Automatic analysis	Section 7
		Subjective assessment	Section 7
		Final comparison	Section 7

Table 5. Review of the methodology

5. CHALLENGES BASED ON BIBLIOGRAPHICAL REVIEW

The bibliographical review has allowed us to identify three fundamental challenges which are dealt with in this section: spotting, synchronisation, and access to the audiovisual content.

Synchronisation is a key feature of both voice-over and off-screen dubbing. Synchronisation is reached thanks to the careful work of audiovisual translators, who rephrase, condense or adapt the text so as to match the images and the time slots available. Moreover, to facilitate the recording by the voice talent, time codes are also included in their script, a task called spotting. Should MT be implemented in the working flow, a specificity would be that translators (or post-editors) would not only correct possible MT errors, but also adapt the text so as to comply with the various types of synchronies (Orero 2006). Ideally, this would require a PE software which displays the audiovisual content and not only the written text.

5.1. SPOTTING

Spotting, also called timing or cueing, is the process of defining in and sometimes out time codes of each voice-over or off-screen dubbing unit. As stated by Díaz-Cintas and Remael (2007: 94), time codes are an essential tool, not only for subtitling, but also for the rest of AVT modes such as dubbing and voice-over. Spotting can be done by an audiovisual translator or by another professional, as it is also the case in subtitling (Sánchez 2004), either before or after the translation. Various scenarios can be found in the profession: (1) the translator is given an already created spotting list, which is the case of templates (Sánchez 2004; Díaz Cintas *et al.* 2007; Kapsaskis 2011; Artegiani *et al*. 2014); (2) the translator is required to do the spotting and decide the time codes; or (3) the translator produces a translation without time codes and another professional does the spotting afterwards. In the second and the third scenarios, the ones considered by Franco *et al.* (2010) in their seminal book on voice-over, it is often the case that translators are given a transcript which includes time codes which do not correspond to the timing of the actual audiovisual content they receive. In the En-Doc corpus, scripts with and without time codes can be found, as illustrated in tables 6 and 7.

25m up in the treetops, old king Zog keeps everything in order...

His kingdom of leaves and branches rises above the Pantanal, the largest wetland in the world, and when the rainy season returns and the floodplains are submerged, his tree becomes a kind of island.

This marsh is so large that the only ones who really know where its boundaries lie are the migrating birds, who leave when it once again becomes dry and yellow.

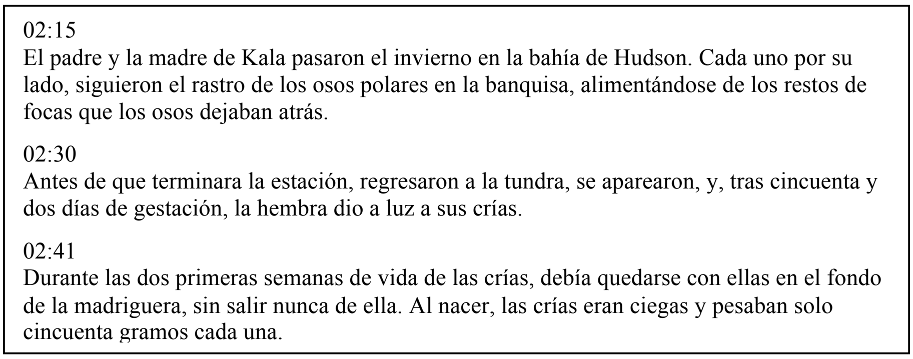
Table 6. Spotting. En-Doc. No Time Codes

02;15 Kala’s father and mother spent the winter on Hudson Bay. Each on its own, they trailed polar bears on the pack ice, feeding on the remains of seals left behind by the bears.

02;28 Before the end of the season, they returned to the tundra, mated and after 52 days of gestation, the female gave birth to her young.

02;43 For the first two weeks of her pups’ lives, she had to stay with them deep in the den without ever coming out. At birth, they were blind and weighed only 50 grams each.

However, all translated scripts in our corpus contain time codes (see table 8), which not always coincide with the time codes in the original script (compare, for instance, the Spanish spotting in table 8 which corresponds to the original in table 7). Thus, translators needed to either introduce the spotting when translating the script or check and rewrite the time codes because they were different.



02:15
El padre y la madre de Kala pasaron el invierno en la bahía de Hudson. Cada uno por su lado, siguieron el rastro de los osos polares en la banquisa, alimentándose de los restos de focas que los osos dejaban atrás.

02:30
Antes de que terminara la estación, regresaron a la tundra, se aparearon, y, tras cincuenta y dos días de gestación, la hembra dio a luz a sus crías.

02:41
Durante las dos primeras semanas de vida de las crías, debía quedarse con ellas en el fondo de la madriguera, sin salir nunca de ella. Al nacer, las crías eran ciegas y pesaban solo cincuenta gramos cada una.

Table 8. Spotting. Spa-Doc. Time Codes

A specificity of voice-over and off-screen dubbing in the corpora and confirmed by the examples in Franco *et al.* (2010) is that, generally, only time codes in (and not out) are included.

An additional difference related to time codes is that in the English original scripts they appear in various formats whilst in the Spanish scripts –for voice-over and off-screen dubbing– the formatting is limited to two. This comes to show that, even in the uncommon scenario in which the time codes in the original script coincide with the target language time codes, adapting their format would be an additional requirement. As summarized in table 9, time codes within En-Doc corpus may indicate minutes and seconds (from type 1 to type 6); hours, minutes and seconds (types 7 to 10); hours, minutes, seconds and frames (from type 11 to 13) or feet (type 14). However, type 6 is the most commonly found among them. In the corpus Spa-Doc only two different time code formats are found: 00:01 (type 6) and 00.01 (type 5), the former being the most common one.

Type	Time code		Type	Time Code
1	(00.02)		8	01:00:10
2	01 08		9	10 04.06
3	0304		10	10.00.03
4	00;04		11	01:00:22:27
5	00.06		12	10 00 07 00
6	00:19		13	(01:08:18:00)
7	00.00.08		14	6.5

Table 9. Types of Time Codes Spotting

All in all, spotting is a must before a documentary is recorded. If MT with post-editing is implemented, dealing with the spotting might be a challenge, be it because time codes will have to be modified (if available) or included (if they do not appear in the original script). Therefore, introducing or correcting the time codes in the script which will be fed into the MT engine, might be an adequate task to increase PE productivity.

5.2. SYNCHRONIZATION

The spotting or assignation of times codes can facilitate the synchronization of text and the audiovisual content according to the three types of synchronies to be reached when translating documentaries (Orero 2006): kinetic synchrony, action synchrony, and isochrony. These synchronisations can only be achieved by confronting the actual translation to the audiovisual content, and in a scenario in which MT is implemented in the working flow, they may have to be carried out during the post-editing phase. However, some automatic strategies to reduce this load may be considered such as limiting the minimum and maximum number of characters per sentence, as already done, for example, by PET (Post-Editing Tool, see http://www.clg.wlv.ac.uk/projects/PET/), a post-editing research tool designed to help users post-edit and assess both MT output and human translations.

5.3. ACCESS TO AUDIOVISUAL CONTENT

As Franco *et al.* (2010) state, the source text in AVT is the audiovisual product, which is made of images and audio. Scripts or transcripts, *i.e.* written texts, are sometimes provided to help the translator but it is not always the case. When machine translating, however, a written original text is needed, be it in the form of a pre-existing script, transcript, or automatic transcription of the audio. As visuals and audio are not considered in the automatic process, it is of the essence that the MT output is revised during the post-editing phase, not only in terms of language adequacy and fluency, but also in terms of written text-audiovisual content synchronisation. In order to do so, access to the visuals is needed, which, to the best of my knowledge, can only be achieved nowadays by using post-editing software plus video player. Available post-editing software, be it commercial CAT tools or applications for research purposes, do not allow rendering of audiovisual content in their interface. This is the case of PET (Aziz *et al.* 2012), CASMACAT (Ortiz-Martínez *et al.* 2012) or TCTool (Font Llitjós 2004b). Although SUMAT looks into the possible integration of MT with AVT, its platform and infrastructure does not integrate neither image nor audio (Del Pozo *et al.* 2013), which means that when carrying out SUMAT tests, participants had to work with standard subtitling software.

6. CHALLENGES BASED ON EN-DOC AND SPA-DOC CORPORA ANALYSES

This analysis is based on the observation of EN-DOC and SPA-DOC corpus and takes a closer look at some of the linguistic issues which affect either scripts' macro- or micro-structures, or both: variety on script format, register variety within the same script, terminology, errors and inaccuracies in the original script, and lexical problems in the original script.

6.1. VARIETY ON THE SCRIPT FORMA

As Franco *et al.* (2010) explain, original scripts formats provided to audiovisual translators differ substantially. After analysing the macro-structure (information contained within the scripts and how it is presented) of all the compiled scripts in the En-Doc corpus, several types of script layouts have been found. The obvious characteristic shared by all scripts is the transcription of narrations plus other speeches, from experts to spontaneous participants. However, it has been observed throughout the corpus that the transcription can be either included in a table which contains additional information or in a plain text document with nothing else but the time-codes.

When the script layout is presented in a table, narrations, also called commentaries, tend to be included under the heading *commentary*, or *comm*, whilst words from experts or spontaneous participants generally follow the term *sync*. It must be stressed that some scripts contain no differentiation between these two types of speakers, and when they appear together, they usually appear under the heading *audio*, *description*, *sync/comm* or *script*. Another feature of the table-based scripts comprised in the corpus is that time codes are always included, under the heading *time codes*, *time code, timecode* or *TC*. Many of these scripts also contain additional information, referring to elements such as images, music or even the mood of each character when talking, with varying degrees of detail. Two examples can be found in tables 10 and 11. Whilst the former indicates that the visuals correspond to boats on a river with no further details ("River-boats"), the latter describes more precisely what is seen ("Local people dancing & playing instruments. Cuts to landscapes") and gives details as to the music that can be heard ("Siddhi Drumming").

TIME CODE	VISUALS	DIALOGUE/NARRATION
10 00 25	River – boats	In 1998, I left Italy and set off for the heart of Africa, to the Congo basin. The focus of my quest… lowland gorillas.

Table 10. Variety of Scripts - En-Doc 1

Timecode	In-Vision	Music	Sync	Narration
10.00.39	Local people dancing & playing instruments. Cuts to landscapes	10.00.44 Siddhi Drumming OUT		African features and rhythms, low thorny forests and the king of the beasts – all establish where we are –or does it?

Table 11. Variety of Scripts - En-Doc 2

On the other hand, and when the script layout is not presented in a table but in a basic text document, it only contains the transcription of the words with speech turns separated into paragraphs and with time codes at the beginning, if available (see table 12).

01 08 Butterflies are particularly well-known for their beautiful shapes and the splendid colours of their wings…
01 17 Their beauty has made them familiar to humans.
01 29 But butterflies are only part of a large family that we are not well acquainted with, the insects, the largest and most successful family of animals on planet Earth.

Table 12. Variety of Scripts - En-Doc 3

Despite the original English scripts can be presented in many different formats, the variety of script layouts in the case of their Spanish counterparts is not as large. Similarly to the original scripts, the translation of the scripts can be either presented in a table (see table 13) or in a plain text document (see table 14), which is the most common option. The latter option sometimes contains indications of the voice talents concerning the pauses to be made (see the slashes in table 14).

Chyros – TC'S	DECLARACIONES	NARRADOR
00.02.14		Antiguas leyendas de marineros hablan de islas misteriosas que se mueven empujadas por la corriente en un mar de tiempo.
02.26		Pueden aparecer y desaparecer de Nuevo en cualquier punto de la enorme extensión del océano. Y llevan el desastre a cualquiera que se acerque demasiado.

Table 13. Variety of Scripts - Spa-Doc 1

____00.01___
____
NARRADOR:
Éste es el parque nacional de Denali, en Alaska. / Aquí las alturas sobrecogedoras…
____00.13___
____
ESCALADOR:
No veo bien…
____00.15___
____
NARRADOR:
Y las tormentas sub-árticas / son los elementos de la vida y la muerte.
(es-59)

Table 14. Variety of scripts - Spa-Doc 2

A correlation between the original script layout and the audiovisual transfer mode used in the translation can be found. The speeches which are normally introduced by the word *narrator* or by no specific heading in the original script correspond to a disembodied voice that is usually off-screen dubbed. They are generally transferred onto the translated script by indicating *narrador* (narrator) or nothing. The ones that are introduced by a specific proper name in the original script correspond to people talking on screen and are usually voiced-over. This is transferred onto the translated scripts by including the name of the on-screen speaker, a nick-name to identify the person, the symbol *VO* or the heading *declaraciones*. On occasions, a narrator or talking head may speak both on- and off-screen, in which cases, the symbols *sync* or *comm* are generally added to indicate whether they appear on- or off-screen in the original version.

All in all, two obvious but relevant conclusions for the use of MT should be highlighted: on the one hand, not all information contained in the original script is to be included in the translated version, and, on the other, translated script layouts are different from the original ones. This means that, most probably, an adapted translation script or template without all the extra information should be created before feeding the MT engine with it. Additional research is needed on how this additional task would impact the productivity and in which scenarios it would be worth it.

6.2. VARIETY OF REGISTERS WITHIN THE SAME SCRIPT

While VO is used to translate the words of interviewed experts and spontaneous dialogue, generally on camera, off-screen dubbing is mostly used for narrators off-camera. Different speakers can coexist within a same wildlife documentary film, and depending on who is talking and the communicative situation, the register may vary:

1. Third person narrator: as stated by León (1998: 18), "(t)he narrator-presenter plays a very important role in television documentary since his voice and statements to camera are the backbone in the structure of the programme." Narrators present and explain facts with the help of images, and sometimes, the presence of experts in the documentary. Their discourse is usually planned, based on a previously written script. In the corpus, their language is generally formal, although more colloquial or non-standard forms may appear occasionally, so as to engage the audience. See for instance, the rhetorical questions used to address the audience in table 15.

00:05
5 extraordinary stories from the wild.

00:08
But watch out because *there's* a twist. One of them is a *fake* created just to test you. Can you tell fact from fiction? Or will you be Fooled By Nature?

00:23
Nature's fantastic feeders.

Table 15. Variety of registers - En-Doc 1

2. First person narrator: narrators may change from a third person commentary to a first-person in order to interact with other participants or to adopt a more subjective approach, as can be seen in table 16.

00:03 COMM Stephen Fry
Twenty years ago my good friend Douglas Adams spent a year tracking down endangered animals together with the zoologist Mark Carwardine. Now it's my turn.

00:15 COMM Stephen Fry
Mark and I are heading off to find out exactly what happened to those species that he'd seen dangling on the edge of extinction two decades ago.

Table 16. Variety of registers – En-Doc 2

Despite being planned, the language on these instances often contains less formal features, as can also be seen in Table 16. These fragments can be re-voiced using voice-over or off-screen dubbing, depending on the market or client.

3. Expert interviewee: interviewees usually appear on-screen and are normally voiced-over in the translated audiovisual product. They do not normally speak from a written text but reply to the questions posed by the interviewer, bearing in mind that

they are addressing a wider audience. This means that the language used is spontaneous or semi-spontaneous. As Matamala (2009: 115) points out, this implies that standard language is generally used, containing some informal features –typical from oral discourse– such as hesitations, false starts, repetitions or anacolutha, *i.e.* syntactical inconsistencies in a sentence.

4. Spontaneous dialogue: it is normally voiced-over in the Spanish product. It varies in its degree of informality depending on the communicative situation and the speaker's idiosyncrasies: from less informal utterances by a speaker talking to the camera, as if addressing the audience, to more informal dialogue exchanges between participants who are almost unaware that the camera is there. As stated by Matamala (2009: 115), interaction between two people who know each other and who do not directly address the audience are more prone to contain informal language and recurrent hesitations, false starts, repetitions, anacolutha, unfinished sentences, interjections and other oral features.

5. Foreign interviewee: non-native speakers might participate in documentaries as experts. When they appear on screen, they can either speak in English or in their own language. If they talk in English, which is a foreign language for them, their speech may contain errors because of lexical and syntactic interferences, and in some cases, borrowed terms from their mother tongue may appear (see table 17).

> 01:06:11 Alex Saragoza
> The *científicos* were the people who implemented his economic policies. These were the people who wrote the legislation for the passage of laws. These were the people who put together the contracts between the Mexican government and foreign companies and so on. They were elitist, some of them were racist, that is they believed in the notion that the biggest problem that Mexico faced was its backward Indian population.

Table 17. Variety of registers - En-Doc 7

If they talk in their own language, sometimes a translation into English is provided in the scripts, as can be seen in Table 18, where the interviewee talks in Spanish and the English translation is provided in italics:

> 01:03:25 Jesus Vargas
> La revolución es un proceso social que tiene una relación íntima con toda la historia de México del siglo diecinueve.
> *The revolution is a social process intimately related to the history of 19th century Mexico.*

Table 18. Variety of registers - En-Doc 8

To guarantee higher quality levels, MT is normally used with texts using one register. The fact that documentaries tend to combine both formal and informal registers, either planned (based on a written script) or spontaneous, proves more demanding for MT. Additionally, specific features such as some repetitions, hesitations and discourse markers may be more difficult to deal with automatically. Still, when translating documentaries from English into Spanish, it is often the case that many of these features (hesitations, repetitions, etc.) disappear in order to reach voice-over isochrony because informative content is prioritized over expressive features (Orero 2006). As these features are not usually translated and they make MT processing more difficult, an option would be to delete them, either manually or automatically, from the script that will be fed into the MT engine.

6.3. TERMINOLOGY

A relevant feature of wildlife documentary films is the inclusion of specific terminology, which varies depending on the topic of the documentary and the general approach, from more to less specialised. Thus, while a documentary film may deal with fishing, another may approach diseases in animals or show the beautifulness of forests and all the fauna and flora they contain. Even if dealing with the same general topic, every wildlife subfield has its specific terminology which may coexist in the same documentary with terminology from other fields.

6.4. ERRORS AND INACCURACIES IN THE ORIGINAL SCRIPT

As pointed out by Franco *et al.* (2010) and Matamala (2009, 2010), original scripts can contain errors and inaccuracies. Dates, names of places and terminology may be wrong; text may be missing from the written script or may appear in the wrong place. Possible errors and inconsistencies in the scripts would not affect the work produced by MT engines, although they could slow down the post-editing process. However, if scripts were checked before being machine translated, the number of errors and inconsistencies in the MT output could be minimized and translators would not have to deal with them during the post-editing process.

6.5. LINGUISTIC INCONSISTENCIES IN THE ORIGINAL SCRIPT

According to Franco *et al.* (2010: 60), it is not uncommon to find an original script with many linguistic mistakes, poor composition and different ways of spelling the same word; a statement that is also proven in the corpus. In the En-Doc corpus, both spelling (*e.g. though* instead of *thought*) and grammar mistakes (*e.g. worlds* instead of *worlds'*; *this* instead of *these*) have been found, as well as punctuation (*e.g.* interrogation or exclamation marks may appear in the middle of a sentence) and capitalization errors (*e.g.* words without a capital letter may appear after a full stop).

It is also worth stressing that sometimes the script presents the sentences cut into neither non-semantic nor grammatical chunks, as they are fit in different rows (see table 19). When this happens, the semantic and grammatical load of the segments is broken and the MT engine performs worst, as the segment can be split in incoherent syntagmas:

00:25 Listen to the
stories each of us
tells you about
ways of obtaining
Unbelievable
food. Then try to spot the fake from this line-up.

Table 19. Linguistic inconsistencies. En-Doc

As Daems *et al.* (2013) explain, errors in the source text affect the efficiency of MT engines and may influence the quality of the target text even after post-editing. Thus, all the previously described mistakes and segmentation problems inevitably have a bearing on the translation produced by MT engines, and ways to overcome these problems need to be found.

7. CHALLENGES BASED ON THE BILINGUAL CORPUS ANALYSIS

An automatic evaluation of the translations produced by eight MT engines and a human-based analysis of the errors found in the MT output was considered an adequate way to predict the challenges of using MT to translate documentary scripts. The results of both the analyses are presented next.

7.1. AUTOMATIC EVALUATION

BLEU and TER measures were produced to evaluate the 50 sentences translated by the 8 selected MT engines (see table 3). These two measures were chosen as they are the more established among MT researchers at present. On the one hand, and according to Papieni *et al.* (2002), the higher the BLEU score is, the better the MT output. On the other, the lower the TER is, the better the MT output is, as it means that the error rate is low (Snover *et al*. 2006). Table 19 presents BLEU and TER scores for each engine:

MT engine	BLEU	TER
Google Translate	29.32	39.41
Apertium	14.19	27.26
Lucy MT	21.20	33.48
Bing	26.88	43.41
Promt	23.99	38.22
Reverso	18.39	25.93
Systran	12.15	3.11
Yandex	27.48	33.63

Table 20. Automatic evaluation scores

Results presented on table 20 show that the engines could be divided into four groups according to their BLEU scores. The top quartile would be formed by the MT engines with higher scores *Google Translate*, *Yandex* and *Bing* (BLEUs from 26.88 to 29.32). The second quartile would include *Promt* and *Lucy MT* (BLEUs from 23.99 to 21.20). In the third, there would only be *Reverso* (BLEU of 18.39), and in the bottom quartile, there would be *Apertium* and *Systran*, the engines with the lower scores (from 12.15 to 14.19). However, if this categorization was made according to TER scores, results would be divided in four groups. The top quartile would include *Bing*, *Google Translate* and *Promt* (38.22 to 43.41), the middle one would have *Yandex*, *Lucy MT*, *Apertium* and *Reverso* (25.93 to 33.63), and the bottom one would only contain *Systran* (3.11).

The highest BLEU score is reached by *Google Translate*'s engine (29.32 points) and the best TER score is attained by *Bing*'s (43.41 points). BLEU scores do not differ much from scores achieved in other experiments that worked with the same language pair, English > Spanish, within the same translation field of AVT (Nakov 2008; Kohen *et al*. 2006; Kohen *et al*. 2007), as their scores also fluctuated between 23.18 and 35.09. Some of these MT engines achieved better BLEU scores than those presented by the SUMAT project (Bywood 2013) and are only six points below the eTITLE's results (Melero 2006). Nevertheless and as an example, the best results are still far from, the ones reached in Vilar *et al.* (2006), where they presented a BLEU score of 48.6 points when they applied customized MT to subtitling (En <> Spa). It should be taken into account, however, that these results are the first available dealing with documentary film translation and are based on free online engines. Engines created specifically for this domain could, of course, yield better results.

7.2. HUMAN EVALUATION

Human evaluation results do not exactly correlate with automatic measures but are to some extent similar. *Google Translate* is the engine that produces fewer errors (69), followed by *Bing* (78) and *Promt* (82). *Yandex* (102) and *Lucy MT* (114) are the next engines with the fewest errors. The three engines that produce more errors are *Apertium* (151), *Systran* (131), and *Reverso* (129). Thus, if engines were grouped according to their number of errors, the group with the highest scores would include exactly the same engines as in the classifications based on TER and BLEU scores.

Engine	**Accuracy**		**Fluency**		**Total**
	Num.	**%**	**Num.**	**%**	
Google	39	56.52	30	43.48	69
Apertium	87	57.61	64	42.38	151
Lucy MT	59	51.75	55	48.38	114
Bing	30	38.46	48	61.54	78
Promt	45	54.88	37	45.12	82
Reverso	69	53.49	60	46.51	129
Systran	64	48.86	67	51.15	131
Yandex	54	52.94	48	47.06	102
TOTAL	447		409		856

Table 21. Human Evaluation. Accuracy & Fluency

As seen in table 21, the majority of errors produced by *Bing* and *Systran*'s engines are related to fluency, while all the other engines have more errors that regard to accuracy. The difference between accuracy and fluency errors produced by *Systran*, *Lucy MT* and *Yandex* is minimal (less than three points between them).

To provide a more detailed analysis, 22 subcategories were considered (12 dealing with accuracy errors and 10 dealing with fluency mistakes), as listed in table 4. No mistakes were found concerning 6 categories: date and time, unit conversion, entity, diacritic accents, punctuation, and unpaired quote marks or brackets. On the contrary, 16 categories reported mistakes: (a) terminology, (b) overly literal, (c) false friend, (d) should not have been translated, (e) number, (f) mistranslations: non-specified errors, (g) omission, (h) addition, (i) untranslated, (j) capitalization, (k) morphology, (l) part of speech, (m) agreement, (n) word order, (o) function words, and (p) unintelligible. Before presenting the results in table 21, an example of each category is presented:

a) *Terminology*

Original sentence: “Okay, so the next dish is monkey faced eel from Port Baker”
Systran’s translation: “La autorización, así que el plato siguiente es anguila hecha frente mono del panadero del puerto”
Back translation: “The authorization, so the dish next is eel done in front of monkey from baker of the port”
Human translation: “De acuerdo, el próximo plateo es anguila caramono de Port Baker”

b) *Overly literal*

Original sentence: “In a small Ugandan fishing village, nestled along the shores of Lake Victoria, crocodiles have recently killed people”
Reverso’s translation: “En un pequeño ugandés el pueblo de pesca, recostado a lo largo de las orillas del lago Victoria, cocodrilos recientemente ha matado a la gente.”
Back translation: "In a small Ugandan [from Uganda] the fishing village, nestled along the shores of Lake Victoria, crocodiles have recently killed people"
Human translation: “En un pequeño pueblo de pescadores de Uganda enclavado en la orilla del lago Victoria, últimamente los cocodrilos han matado gente.”

c) *False friend*

Original sentence: “Oh, right”
Yandex’s translation: “Oh, a la derecha”
Back translation: “Oh, to the right”
Human translation: “Ah, perfecto”

d) *Should not have been translated*

Original sentence: “Okay, so the next dish is monkey faced eel from Port Baker.”
Apertium’s translation: “Okay, así que el plato próximo es monkey anguila afrontada de Panadero de Puerto.”
Back translation: “Okay, so the dish next is monkey eel faced from Baker of Port.”

Human translation: "De acuerdo, el próximo plato es anguila caramono de Port Baker."

e) *Number*

Original sentence: "My gun won't fire. My gun won't fire"
Yandex's translation: "Mis armas no de fuego. Mis armas no de fuego"
Back translation: "My guns not of fire. My guns not of fire"
Human translation: "La escopeta no dispara. La escopeta no dispara."

f) *Mistranslations: non-specified errors*

Original sentence: "She quietly leaves the group and lies down on a secluded spot to await her delivery"
Lucy's translation: "Silenciosamente deja el grupo y se tumba en una mancha/sitio retirada para esperar a su entrega"
Back translation: "She quietly leaves the group and lies down on a secluded spot [patch/place] to await her delivery"
Human translation: "Abandona silenciosamente el grupo y se tumba en un lugar apartado para esperar el momento del parto"

g) *Omission*

Original sentence: "But he suspected something else was at work as well."
Bing's translation: "Pero sospechaba que algo [missing: más] estaba obrando así."
Back translation: "But he suspected something [missing: else] was at work as well."
Human translation: "Pero sospechaba que había algo más."

h) *Addition*

Original sentence: "This is better with garlic."
Systran's translation: "Esto es mejor con el ajo."
Back translation: "This is better with the garlic."
Human translation: "Están más buenos con ajo."

i) *Untranslated*

Original sentence: "Then a group of killer whales headed towards shore, as if they intended to strand"
Apertium's translation: "Entonces un grupo de killer las ballenas encabezadas hacia shore, cuando si pretendieron a strand."
Back translation: "Then a group of *killer* the whales headed [meaning "led"] towards *shore*, when if intended to *strand*."
Human translation: "Entonces un grupo de orcas se dirigió hacia la orilla, como si quisieran quedarse varadas."

j) *Capitalization*

Original sentence: “He’s dominated the prairie for some years now, and few have dared comfort him face to face.”
Promt’s translation: “Ha dominado la Pradera durante algunos años ahora, y pocos se han atrevido a oponerse a él cara a cara.”
Back translation: “He's dominated the Prairie for some years now, and few have dared comfort him face to face.”
Human translation: “Ya hace algunos años que domina la llanura y pocos se han atrevido a enfrentarse a él cara a cara.”

k) *Morphology*

Original sentence: “Between the people, the pavement, and the most overprotective laws in the country”
Lucy’s translation: “Entre la gente, la acera, y las leyes más sobreproteccionistas del país.”
Back translation: “Between the people, the pavement, and the most overprotectionist laws in the country”
Human translation: “Entre la gente, el pavimento, y estas leyes tan sobreprotectoras del país”

l) *Part of speech*

Original sentence: “Her body strength is recovering quickly, and her calf now kicking”
Google’s translation: “Su fuerza del cuerpo se está recuperando, y su cría ya patadas”
Back translation: “Her strength of the body is recovering, and her calf already kick [noun]”
Human translation: “Está recuperando las fuerzas rápidamente y la cría ya le da patadas”

m) *Agreement*

Original sentence: “It’s surprising crocs would spend so much energy climbing up this cliff”
Bing’s translation: “Es sorprendentes crocs pasaría tanta energía subiendo este acantilado.”
Back translation: “It's surprising [plural] crocs would spend [singular] so much energy climbing up this cliff.”
Human translation: “Es increíble que los cocodrilos gasten tanta energía subiendo por este acantilado.”

n) *Word order*

Original sentence: “Her body strength is recovering quickly, and her calf now kicking.”
Systran’s translation: “Su fuerza del cuerpo se recupera rápidamente, y su becerro ahora dando patadas”

Back translation: "Her strength of body is recovering quickly, and her calf now kicking."
Human translation: “Está recuperando las fuerzas rápidamente y la cría le da patadas.”

o) *Function words*

Original sentence: “I feel that it’s so important for me to try to get the Toga people understand what we have in our own back yard is something very unique.”
Google’s translation: “Siento que es tan importante para mí tratar de conseguir [que] la gente Toga entienden [que] lo que tenemos en nuestro propio patio trasero es algo muy especial”
Comment: In Spanish it is to introduce function words that are not used or necessary in English.
Human translation: “Es muy importante que haga entender a los tonganos que lo que tenemos aquí es algo único.”

p) *Unintelligible*

Original sentence: “If it’s swimming towards you, get it over the entire head and tighten it up.”
Systran’s translation: “Si esto nada hacia usted, conseguirlo sobre la cabeza entera y apretarlo encima de.”
Back translation: “If this swims towards you, get it [achieve it] over the entire head and tighten it up above.”
Human translation: “Si nada hacia vosotros, la metéis por la cabeza y tensáis.”

As shown in table 22, the categories with most errors are (m) *agreement* with 186 cases, (f) *mistranslations: other* with 133, and (i) *untranslated* with 86. While the majority of errors in *Google Translate* and *Apertium* are *untranslated* and *mistranslations: other*, all the others engines deal mostly with problems regarding *agreement*. The categories following the lead are (b) *overly literal* with 77 errors, (n) *word order* with 72 and (a) *terminology* with 63. In the central part of the table there are the categories (l) *part of speech* with 56 errors, (g) *omission* with 44, (p) *unintelligible* with 42, (o) *function words* with 39 and (h) *addition* with 37. The categories with lower errors are (j) *capitalization* with 13 errors and (c) *false friends* with 5, as well as three categories with a single error: (d) *should not have been translated*, (e) *number* and (k) *morphology*.

Engine	a	b	c	d	e	f	g	h	i	J	k	l	M	n	o	p	TOTAL
Google	5	4	0	0	0	15	2	8	5	0	0	7	13	5	3	2	69
Apertium	11	12	0	0	0	14	5	1	44	0	0	13	23	16	4	8	151
Lucy MT	12	9	0	0	0	21	6	1	10	0	1	6	28	10	4	6	114
Bing	7	1	0	0	0	10	8	1	3	1	0	6	25	5	7	4	78
Promt	7	13	2	0	0	11	3	2	7	2	0	1	18	6	4	6	82
Reverso	7	16	3	0	0	21	9	8	5	5	0	3	28	9	5	10	129
Systran	9	11	0	0	0	25	7	6	6	4	0	3	32	14	9	5	131
Yandex	5	11	0	1	1	16	4	10	6	1	0	17	19	7	3	1	102
TOTAL	63	77	5	1	1	133	44	37	86	13	1	56	186	72	39	42	859

Table 22. Human evaluation. Types of errors

To sum up, human evaluation results give us an indication of the most frequent type of mistakes audiovisual translators would have to correct in a post-editing phase: agreement, mistranslated, and untranslated words. Additionally, it indicates that, from the freely available online engines in the English > Spanish combination, *Google Translate* appears to be the best MT engine, followed by *Bing* and *Promt* at least for this study's sample excerpts from documentaries. Although this data may not be relevant for a company deciding to develop their own MT system, (as the analysis is only based on 50 segments and companies normally rely on internal systems specifically developed to satisfy their needs) it is a first step in an underexplored area that might be useful for other scenarios, such as journalistic translation, in which online software can be used.

8. DISCUSSION: POSSIBLE SOLUTIONS

The bibliographical review and the three corpora analysis have shown several challenges that would have to be addressed in order to integrate MT into the translation process of wildlife documentary films. Before presenting a new workflow to help overcome the challenges, some solutions are proposed for each of the above mentioned challenges.

First of all, solutions regarding the challenges encountered in the bibliographical review –spotting, synchronization and access to the audiovisual content– will be presented. In professional practice, audiovisual translators usually synchronize the visuals and their translation, and are sometimes required to do the spotting, *i.e.* to include the time codes. If MT was to be included in the process of translating documentaries, the MT output would not only have to be corrected during the post-editing stage, but also revised to comply with the various types of synchronies at stake. Correct time codes would also have to be included during the post-editing. In order to do so, full access to the visual content would be required. A suggested scenario to solve these issues would be to include a pre-editing phase (Volk 2009; Gerlach *et al.* 2013) in which a time-coded script to be used by translators working into different languages would be created, and additionally, it would be necessary that PE software includes a video player. A tool to limit the maximum number of characters or words per sentence could be also helpful, like PET does for subtitling, as it could help post-editors know how much space they have for each voice-over or off-screen dubbed unit.

Secondly, solutions to the issues found in the analysis of the corpora are proposed. According to the analysis, there are many types of script layouts in English, and to a lesser extent, in Spanish. Therefore, standardizing the script layouts in the original language seems a field in which further work needs to be done. In the meanwhile, creating an MT friendly template every time a documentary is to be translated seems to be a possible solution. This template would contain plain text (not tables) and would be created, again, in a pre-editing phase, ideally with automatic tools that extract the original dialogue from the audiovisual product. It remains to be seen whether this proposed scenario would be feasible when the original documentary is to be translated into one single language or would rather be used in multilingual contexts. Researching this aspect, though, is beyond the scope of this paper.

As for the mixing of various language registers in the same audiovisual programme, a possible solution could be to create a domain-specific engine with wildlife documentaries. Although register-related problems would persist, terminological and lexical problems would hypothetically decrease and reduce the post-editors workload. In order to minimize register challenges, features such as hesitations or repetitions could be erased from the scripts in the pre-editing phase before feeding them into this domain-specific engine.

As for linguistic inconsistencies and errors, they could be rectified either in pre- or post-editing. On the one hand, spelling mistakes and other linguistic problems due to original text formatting could be pre-edited, as they might influence the quality of the MT output. On the other hand, capitalizations and other types of linguistic inconsistencies and errors could be solved during post-editing, as they do not have an impact on the output. Nevertheless, correcting them in the pre-editing phase would be better, as the MT output would drag almost no errors from the original script. In this way post-editors could focus mainly on correcting linguistic errors produced by the MT engine (mainly agreement mistakes and mistranslation, according to our analysis) and solving problems regarding domain-specific issues.

All in all, the analysis has shown that there are problems broadly found in MT which are generally solved through post-editing, but there are also specific challenges related to this text type and audiovisual modality which may be better dealt with in an additional pre-editing phase. What remains to be seen is the impact of this phase in the whole process in terms of time and productivity. However, the availability of a script specifically prepared for MT would have two clear implications. On the one hand, the same script could be used when translating into a different language. On the other, it could let post-editors concentrate more on voice-over and off-screen dubbing specific features. Thus, the following workflow, divided in three steps, is proposed in table 23:

Phase	Tasks
Before translating	1. Build a domain-specific MT engine for wildlife documentary scripts
Pre-editing	1. Spotting 2. Creation of an MT-friendly template 3. Elimination of linguistic inaccuracies 4. Elimination of specific features such as hesitations, repetitions and fluffs.
Machine Translating	1. Machine translate the template
Post-editing	1. Check synchronization between text, images and sound 2. Check register 3. Check terminology 4. Check grammatical and syntactical errors and inaccuracies 5. Solve linguistic inconsistencies especially in terms of accuracy and fluency In order to do so more efficiently, a PE tool including a video display and tool to count words should be used.

Table 23. Possible solutions. Workflow

9. CONCLUSIONS AND FURTHER RESEARCH

In conclusion, this article has presented the results of a corpus analysis which has allowed us to identify the main challenges that using MT for the translation of wildlife documentaries might pose: spotting, synchronization, access to audiovisual content, variety on the script format, register variety within a same script, terminology, errors and inaccuracies in the original script, linguistic inconsistencies in the original script, and typical errors in the machine translated output. Three solutions have been proposed to increase the efficiency of post-editing machine translated wildlife documentaries: firstly, pre-editing, as it has been considered to be the answer to challenges such as the inclusion of time-codes, the elimination of certain problematic features (repetitions, hesitations, etc.), and the revision of language of content-related mistakes. Pre-editing has been proposed as a potential solution as it would allow for faster post-editing, an aspect already proven in other contexts such as user-generated content translation (Sertan *et al*. 2014). Secondly, building a domain-specific engine has been proposed as a possible solution to deal with specific terminology, and thirdly, working with templates has been considered a possible strategy when dealing with a large variety of script formats. Furthermore, the analysis has pointed out the relevance of having access to the audiovisual material, as without it, no successful spotting or synchronization could be made. However, the lack of PE software that allows the inclusion of audiovisual content is still a technical challenge to be overcome. Were all these proposed solutions implemented, post-editing would probably be more efficient and would allow translators to focus on the most specific aspect of this translation mode: synchronisation. Therefore, taking into account the specificities of the genre and the layout characteristics of the scripts, a combination of pre- and post-editing seems to be the most feasible scenario if MT is included in the process of translating wildlife documentary films. Still, further research to prove this hypothesis and its impact on the final workflow needs to be carried out.

Additionally, the analysis has considered a scenario in which a specific engine cannot be built and free online software is used. The analysis of a corpus of machine translated wildlife documentary excerpts has allowed us to identify the main mistakes produced by free online MT engines, namely agreement, mistranslated and untranslated words. This analysis has also shown that, even when using non-specific MT engines, the results of the automatic quality measures are similar to those achieved in other relevant experiments with the same language pair. Such results seem to indicate that future research can be promising as there is still much room for improvement by using, for instance, domain specific MT. Moreover, as many mistakes found in the analysis are of a repetitive nature, and the use of automatic systems to constrain propagation could speed-up the PE task.

To sum up, both the results of the analysis and the presented challenges and solutions seem to indicate that further research on the inclusion of MT in the process of translating wildlife documentaries is advisable. Future investigations could include a similar analysis with other language pairs and translation engines, as well as an analysis of the post-editing effort compared to the human translation effort in which both objective measures and subjective data could be obtained. This future study could also consider other variables such as the inclusion or non-inclusion of a pre-editing phase. All in all, the MT of wildlife documentaries is a novel topic which opens new research opportunities to which I have tried to contribute by carrying out this exploratory research.

BIBLIOGRAPHICAL REFERENCES

Almeida, Giselle de and Sharon O'Brien. "Analysing Post-Editing Performance: Correlations with Years of Translation Experience." *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. Ed. EAMT. May 2010. 27-28. Available at: http://www.mt-archive.info/EAMT2010-Almeida.pdf [Consulted: 29-June-2016].

Armstrong, Stephen *et al.* "Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation." *Translating and the Computer*. Ed. ASLIB28. November 2006. 13 pp. Available at: http://www.mt-archive.info/Aslib-2006-Armstrong.pdf [Consulted: 29-June-2016].

Artegiani, Irene and Dionysios Kapsaskis. "Template files: Asset or Anathema? A Qualitative Analysis of the Subtitles of *The Sopranos*." *Perspectives* 22.3 (2014): 419-436.

Aziz, Wilker *et al.* "PET: a Tool for Post-Editing and Assessing Machine Translation." *The 8th International Conference on Language Resources and Evaluation (LREC2012)*. Ed. LREC'12. May 2012. 3982-3987. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf [Consulted: 29-June 2016].

Baldry, Anthony and Paul J. Thibault. *Multimodal Transcription and Text Analysis. A Multimedia Toolkit and Coursebook*. London & Oakville, UK: Equinox, 2006. 270 pp.

Bywood, Lindsay *et al.* "Parallel Subtitle Corpora and their Applications in Machine Translation and Translatology." *Perspectives: Studies in Translatology*. *Special Issue: Corpus Linguistics and Audiovisual Translation: in Search of a Integrated Approach* 21. 4 (2013): 595-610.

Daems, Joke *et al.* "On the Origin of Errors: A Fine-Grained Analysis of MT and PE Errors and their Relationship." *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC2014)*. Ed. LREC'14. May 2014. 62-66. Available at: http://www.lrec-conf.org/proceedings/ lrec2014/pdf/532_Paper.pdf [Consulted: 29-June2016].

Díaz Cintas, Jorge and Pilar Orero. "Screen Translation, Voice-over." *Encyclopedia of Languages*. Ed. Keith Brown. London: Elsevier, 2005. 473-475.

Díaz Cintas, Jorge and Aline Remael. *Audiovisual Translation: Subtitling*. Manchester: St. Jerome, 2007.

Dijk, Teun A. van "A Note on Linguistic Macro-Structures." *Linguistische Perspektiven*. Ed. Abraham P. Ten Gate and Peter Jordens. Tübingen: Niemeyer, 1973. 75-87.

Dijk, Teun A. van and Walter Kintsch. *Strategies of Discourse Comprehension*. New York: Academic Press, 1983.

Eckersley, Helen. "Systems for Evaluating Translation Quality." *Multilingual Computing and Technology* 3.3 (2002): 39-42.

Espasa, Eva. "Myths about Documentary Translation." *Topics in Audiovisual Translation*. Ed. Pilar Orero. Amsterdam: John Benjamins, 2004. 183-197.

Fernández, Anna *et al.* Enhancing Sensorial and Linguistic Accessiblity with Technology: Further Developments in the TECNACC and ALST projects. Paper presented at the *VI Media4All Conference*. Dubrovnik, 25-27 September 2013.

Fishel, Mark *et al.* "From Subtitles to Parallel Corpora." *Proceedings of the 16th Annual Conference of the European Association for Machine Translation EAMT 2012*. Ed: Cettolo, Mauro *et al.* May 2013. 3-6. Available at: http://hnk.ffzg.hr/bibl/eamt2012/EAMT-2012.pdf [Consulted: 29-June-2016].

Font Llitjós, Ariadna *et al.* "Error Analysis of Two Types of Grammar for the Purpose of Automatic Rule Refinement." *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*. Ed. AMTA. 2004a. 187-196. Available at: http://works.bepress.com/ jaim e_ca rbonell/10/ [Consulted: 29-June-2016].

Font Llitjós, Ariadna and Jaime G. Carbonell. "The Translation Correction Tool: English-Spanish User Studies." *Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC2014).* LREC'04. 2004b. 4 pp. Available at: http://repository.cmu.edu/cgi/viewcontent.cgi? article=1430& context=isr [Consulted: 29-June-2016].

Franco, Eliana P.C. "Documentary Film Translation: A Specific Practice?" *Translation in Context: Selected Contributions from the EST Congress, Granada 1998*. Ed. Andrew Chesterman *et al*. Amsterdam: John Benjamins, 2000. 233-242.

—. "Inevitable Exoticism: The Translation of Culture-Specific Items in Documentaries." *La traducción en los medios audiovisuales*. Ed. Frederic Chaume and Rosa Agost. Estudios sobre la Traducció, vol. 7. Castelló de la Plana: Publicacions de la Universitat Jaume I, 2001a. 177-181.

—. "Voiced-over Television Documentaries: Terminological and Conceptual Issues for their Research." *Target* 13.2 (2001b): 289-304.

Franco, Eliana *et al*. *Voice-over Translation: an Overview*. Bern: Peter Lang, 2010.

Freitag, Markus *et al.* "EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project." *Proceedings of the 10th International Workshop for Spoken Language Translation (IWSLT 2013).* IWSLT'13. 2013. 8 pp. Available at: http://workshop2013.iwslt.org/downloads/EUBRIDGE_MT_Text_Translation_of_Talks_in_the_EUBRIDGE_Project.pdf [Consulted: 29-June-2016].

García Luque, Francisca. "De cómo 'domesticar' un documental de divulgación científica en el proceso de traducción. Estudio de la versión en español de *L'Odyssée de l'espèce*." *SENDEBAR* 22 (2011): 235-263.

Georgakopoulou, Yota. "Challenges for the Audiovisual Industry in the Digital Age: Accessibility and Multilingualism." *MetaNet Forum*. Ed. MetaNet. 2010. 1 p. Available at: http://www.meta-net.eu/events/meta-forum-2010/slides/META-FORUM2010_Georgakopoulou.pdf [Consulted: 29th June 2016].

Gerlach, Johanna *et al.* "Combining Pre-editing and Post-editing to Improve SMT of User-Generated Content." *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*. Eds. EAMT. September 2013. 45-53 Available at: http://www.mt-archive.info/10/MTS-2013-W2-Gerlach.pdf [Consulted: 29-June-2016].

Hurtado Albir, Amparo. *Traducción y traductología*. Madrid: Cátedra, 2001.

Hurtado Albir, Amparo. "Compentence-Based Curriculum Design for Training Translators." *The Interpreter and Translator Trainer* 1.2. (2007): 163-195.

Hutchins, William John. *Machine Translation: Past, Present, Future*. Ellis Horwood Series in Computers and their Applications. Chichester, UK: Ellis Horwood, 1986.

Joscelyne, Andrew. *LSPs in the MT Loop: Current Practices, Future Requirements*. TAUS Report. 2009. Available at: <https://www.taus.net/think-tank/reports/translate-reports/lsps-in-the-mt-loop-current-practices-future-requirements [Consulted: 29-June-2016].

Kapsaskis, Dionysios. "Professional Identity and Training of Translators in the Context of Globalisation: the Example of Subtitling." *The Journal of Specialised Translation* 16 (2011): 162-184.

Kohen, Philipp and Christov Monz. "Manual and Automatic Evaluation of Machine Translation between European Languages." *Proceedings of the Workshop on Statistical Machine Translation*. Ed. ACL. 2006. 102-212. Available at: http://homepages.inf.ed.ac.uk/pkoehn/publications/shared-task-wmt2006.pdf [Consulted: 29-June-2016].

Kohen, Philipp and Hieu Hoang. "Factored Translation Models." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Ed.: ACL. 2007. 868-876. Available at: http://homepages.inf.ed.ac.uk/pkoehn/publications/emnlp2007-factored.pdf [Consulted: 29-June-2016].

Koponen, Maarit. "Correctness of Machine Translation: a Machine Translation Post-editing Task." Paper presented at the *3rd MOLTO Project Meeting*. (Helsinki, 2011). Available at: http://www.molto-project.eu/sites/default/files/mol to_2011 0902_mkoponen.pdf [Consulted: 29-June-2016].

Läubli, Samuel *et al.* "Combining Statistical Machine Translation and Translation Memories with Domain Adaptation." *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Ed: NEALT. Proceedings Series, vol. 16. Oslo: Oslo University, 2013. 9 pp. Available at: http://stp.lingfil.uu.se/nodalida/2013/pdf/NODALIDA30.pdf [Consulted: 29th-June-2016].

León, Bienvenido. "Science Popularisation through Television Documentary: A Study of the Work of British Wildlife Filmmaker David Attenborough." *5th International Conference of Science and Technology*. Ed. Nigel Sanitt. Berlin: The Pantaneo Forum, 1998: 17-19.

Lommel, Arle. *Multidimensional Quality Metrics*. Paper presented at the *META-FORUM 2013*. (Berlin 2013). Available at: http://www.meta-net.eu/events/meta-forum-2013/talks/arlelommel.pdf [Consulted: 29-June-2016].

Matamala, Anna. *La traducción para voice-over: Online Module for the Master's Degree in Audiovisual Translation*. Barcelona: Universitat Autònoma de Barcelona, 2002.

—. "Teaching Voice-over Translation." *Languages and the Media: New Markets, New Tools. Conference Proceedings*. Ed. ICWE. Berlin: ICWE, 2004. 24-26.

—. “Teaching Voice-over Translation: A Practical Approach.” *The Didactics of Audiovisual Translation*. Ed. Jorge Díaz Cintas. Amsterdam: John Benjamins, 2008. 231-262.

—. “Main Challenges in the Translation of Documentaries.” *New Trends in Audiovisual Translation*. Ed. Jorge Díaz Cintas. Bristol: Multilingual Matters, 2009a. 109-120.

—. “Translating Documentaries: from Neanderthals to the *Supernanny*.” *Perspectives: Studies in Translatology* 17.2 (2009b): 93-107.

—. “Terminological Challenges in the Translation of Science Documentaries: a Case-Study.” *Across Languages and Cultures* 11.2 (2010): 255-272.

Matamala, Anna and Carla Ortiz-Boix. *Accessibility and Multilingalism: An Exploratory Study on the Machine Translation of Audio Descriptions*. Forthcoming.

Melero, Melero *et al.* “Automatic Multilingual Subtitling in the eTITLE Project.” *Proceedings of ASLIB Translating and the Computer 28*. Ed: ASLIB. 2006. 18 pp. Available at: http://citeseerx.ist.psu.edu/viewdoc/download ?doi=10.1.1.107.6011&rep=rep1&type=pdf [Consulted: 29-June-2016].

Nakov, Preslav. “Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Parapharsing, Tokenization and Recasing.” *Proceedings of the Third Workshop on Statistical Machine Translation*. Ed. ACL. 2008. 147-150. Available at: http://dl.acm.org/citation.cfm?id=1626414 [Consulted: 29-June-2016].

Nord, Christiane. “Text Analysis in Translator Training.” *Teaching Translation and Interpreting: Training, Talent, and Experience*. Ed. Cay Dollerup and Anne Loddegaard. Amsterdam: John Benjamins, 1991. 39-47.

—. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St Jerome Publishing, 2014.

O’Brien, Sharon. “Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output.” *Across Languages and Cultures* 7.1 (2006): 1-21.

Orero, Pilar. “The Pretended Easiness of Voice-over Translation of TV.” *The Journal of Specialized Translation* 2 (2004): 76-96.

—. “Synchronization in Voice-Over.” *Aspects of Translation*. Ed. José María Bravo. Valladolid: Universidad de Valladolid, 2006. 255-264.

—. “Voice-over: A Case of Hyper-Reality.” *EU High Level Scientific Conference Series. MUTRA Proceedings.*. Ed. MuTra. 2007. 9 pp. Available at: http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Orero_Pilar.pdf [Consulted: 29-June-2016].

Ortiz-Boix, Carla. *Tecnologies per a l'audiodescripció: estudi sobre l'aplicació de la traducció automàtica i la síntesi de parla a l'audiodescripció en castellà*. Master Thesis. Universitat Autònoma de Barcelona, 2012.

Ortiz-Martínez, Daniel *et al.* "The CASMACAT Project: The Next Generation Translator's Workbench." *Proceedings of the 7th* Jornadas en tecnologia del Habla *and the 3rd Iberian SLTech Workshop (IberSPEECH)*. Ed. ATVS *et al.* 2012. 326-334. Available at: http://www.casmacat.eu/uploads/Main/iberspeech1.pdf [Consulted: 29-June -2016].

Pönniö, Kaarina. "Voice over, narration et commentaire." *Communication audiovisuelle et transferts linguistiques/Audiovisual Communication and Language Transfer*. Ed. Yves Gambier. Special issue of *Translatio* (FIT Newsletter / Nouvelles de la FIT). Strasbourg: Fédération Internationele des Traducteurs (FIT), 1995. 303-307.

Pozo, Arantza del *et al. SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report*. Ed.: Pozo, Arantza del. 2012. Available at: <http://cordis.europa.eu/fp7/ict/language-technologies/docs/sumat-annual-report-2012.pdf [Consulted: 29-June-2016].

Ray, Rebecca, ed. *LISA Best Practice Guides. Implementing Machine Translation*. .2004. 74 pp. Available at: http://www.translationoptimization.com/papers/DillingerLommel_MT_BPG.pdf [Consulted: 29-June.2016].

Remael, Alice. "Audiovisual Translation." *Handbook of Translation Studies*, vol. 1. Ed. Yves Gambier, Yves and Luc van Doorslaer. Amsterdam: John Benjamins, 2010. 12-17.

Sánchez, Diana. "Subtitling Methods and Team-Translation." *Topics in Audiovisual Translation*, vol 56. Ed. Pilar Orero. Amsterdam/Philadelphia: John Benjamins, 2004. 9-17.

Sertan, Violeta *et al.* "A Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation." *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC2014)*. Ed. LREC'14. May 2014. 1793-1799. Available at: http://www.lrec-conf.org/proce edings/lrec2014/pdf/676_Paper.pdf [Consulted: 29-June-2016].

Sousa, Sheila de *et al.* "Assessing the Post-editing Effort for Automatic and Semi-automatic Translations of DVD subtitles." *Proceedings of Recent Advances in Natural Language Processing*. Ed. ACL (2011): 97-103. Available at: http://aclweb.org/anthology/R11-1014 [Consulted: 29-June-2016].

Uszkoreit, Hans and Arle Lommel. *Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment*. Ed. QT Launch Pad Project. 2013. Available at: http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf [Consulted: 29-June-2016].

Vilar, David *et al.* "Error Analysis of Statistical Machine Translation Output." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Ed. LREC. 2006. 697-702. Available at: http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf [Consulted: 29-June-2016].

Volk, Martin. "The Automatic Translation of Film Subtitles. A Machine Translation Success Story?" *Journal for Language Technology and Computational Linguistics* 23.2 (2008): 113-125.

Volk, Martin *et al.* "Machine Translation of TV Subtitles for Large Scale Production." *Proceedings of the Second Joint EM+ICNGL Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC '10)*. Ed. Zhechev, Ventsislav *et al.* 2010. 53-62. Available at: http://www.mt-archive.info/10/JEC-2010-Volk.pdf [Consulted: 29-June-2016].

Waibel, Alex. *EU-Bridge Newsletter. 1st Edition.* (2012): 8 pp. Available at: http://project.eu-bridge.eu/img/Newsletter_1_edition.pdf [Consulted: 29-June-2016].

Williams, Malcolm. "The Application of Argumentation Theory to Translation Quality Assessment." *META* 46.2 (2001): 327-344.

FILM REFERENCES

Planet Earth, Alistair Fothergill. Producer: BBC Natural History, UK 2006.

The Crocodile Hunter, John Staiton. Producer: The Best Picture Show Company, Australia 1997-2004.

A2. 2nd Article

The Journal of Specialised Translation Issue 26 – July 2016

Post-editing wildlife documentary films: A new possible scenario?

Carla Ortiz-Boix, Universitat Autònoma de Barcelona
Anna Matamala, Universitat Autònoma de Barcelona

ABSTRACT

Several studies have proven that, when machine translation followed by post-editing is used to translate general and specialised texts, there is an increase in the productivity, as the post-editing effort is lower than translating *ex novo*. Although the use of machine translation and post-editing has been investigated in Audiovisual Translation, this has never been researched in non-fictional audiovisual genres in which voice-over and off-screen dubbing are applied. Using an English wildlife documentary film as the source text, and Spanish as the target language, this study intends to research whether post-editing involves more or less effort than translating a documentary. Conclusions on the experiment described in this article, in which 12 Audiovisual Translation MA students took part, seem to indicate that post-editing involves less effort than translating.

KEYWORDS

Audiovisual translation, machine translation, post-editing, voice-over and off-screen dubbing.

1. Introduction

In the last two decades, the use of Machine Translation (MT) followed by post-editing when applied to general and specialised translation has been expanding. Such growth has affected not only the market (TAUS, 2009), but also research on post-editing. However, the market of audiovisual translation has barely been affected. Research studies that intend to include MT and post-editing into the process of translating audiovisual products only started a few years ago thanks to European projects such as eTITLE (Melero *et al*, 2006) or, more recently, SUMAT (Del Pozo *et al*, 2013), both focusing on subtitling. The promising results presented by the latter led us to believe that applying MT and post-editing to other audiovisual translation modalities might be feasible and worth researching. This has been precisely the aim of the ALST project (Matamala *et al*, 2012): to investigate the possible application of MT and post-editing into two oral audiovisual transfer modes, namely audio description and voice-over.

The research presented in this article is part of the aforementioned ALST project (FFI-2012-31024), which is financed by the Spanish "Ministerio de Economía y Competitividad", and focuses exclusively on wildlife documentary films which are translated by means of voice-over and off-screen dubbing. Voice-over is the revoicing of an audiovisual text in another language in which a translating voice is superimposed on the original voice (Franco *et al,* 2010). It is frequently used in non-fictional audiovisual genres, especially when speakers appear on-screen, but also in fictional TV programmes in Eastern Europe. On the other hand, off-screen dubbing generally refers to the audiovisual transfer mode used to revoice off-screen narrations in which the original voice is substituted by a target language version (Franco *et al,* 2010). Wildlife documentary films have been selected because, according to a preliminary study by Ortiz-Boix (forthcoming) on a corpus of documentaries, many elements (such as the promising results of the analysed free online MT engines, and the types of errors these engines produce) seem to indicate that it would be feasible to apply MT to this specific genre. However, testing this new scenario in comparison with existing practices with users is yet to be carried out. This is precisely the aim of the research described in this paper: to compare the effort when post-editing a machine translated wildlife documentary and when translating it. Our hypothesis is that post-editing will require less effort than translating.

The article is structured as follows: Section 2 discusses the theoretical approach taken in this paper. In section 3, the methodology used is explained, describing in detail the experiments carried out in June 2014, as well as the methods used to analyse the data. Section 4 discusses the results, taking into account the different types of efforts analysed (temporal, technical, cognitive), and section 5 presents the conclusions and avenues for further research.

2. Theoretical approach: post-editing effort in audiovisual translation

This section defines post-editing and how the effort involved has been measured in previous experiments. It also highlights the specificities of the audiovisual transfer modes under analysis.

Post-editing is the "term used for the correction of MT output by human linguists/editors" (Veale and Way, 1997, cited in O'Brien, 2010:1) and, therefore, "the task of the post-editor is to edit, modify and/or correct pre-translated text" (Allen, 2003:297). Post-editing can basically be carried out on two different levels: minimal or light, and full (Allen, 2003:304-306) and, depending on the level of post-editing used, the required effort will vary.

During the last decade, defining and measuring effort within post-editing research has been in the spotlight, thanks to works carried out by Krings (2001), O'Brien (2004, 2005 and 2006) or Martínez (2003), to name just a few. Krings (2001) led the way by determining how to calculate such effort and setting the standard for the majority of the other works on this topic. According to Krings (2001), post-editing effort can be divided into three types: temporal, technical and cognitive. Temporal effort is understood as the time taken to post-edit a document. Technical effort refers to the number of keystrokes, mouse movements and clicks. And cognitive effort applies to "the extent and type of cognitive processes that

must be activated to remedy a deficiency in the MT output" (Krings 2001:179).
While temporal and technical efforts can be directly observed thanks to keylogging software, as can be seen in Allen (2001), Martínez (2003) or Tatsumi and Roturier (2010), cognitive effort cannot be directly observed. Hence, several methods have been used to observe it: Krings (2001) used Think-Aloud Protocols, although he later realised that verbalising all the movements slowed down the process. O'Brien (2004) observed cognitive effort using Translog, a keylogging software. Although Translog did not permit the direct observation of cognitive effort, it did succeed in measuring the number, location and duration of pauses, which were all considered good indicators of cognitive load (O'Brien 2006; Shreve *et al* 2011). Eye-tracking, a non-intrusive equipment that records eye movements and fixations, is another tool used to measure cognitive effort (O'Brien, 2011). To determine the cognitive load of post-editing effort, processing speed, average fixation time and count are generally taken into account. More recently, Lacruz *et al* (2014a; 2014b) have claimed that there are two formulae that correlate well with cognitive effort: average pause ratio (APR) and pause to word ratio (PWR). According to them, a low APR (the least possible amount of time spent pausing) combined with a high PWR (the most possible time spent pausing per word) are associated with high levels of cognitive effort. To allow for a lower level of applied cognitive effort, a combination of high APR and low PWR, would be beneficial. Both data can be obtained using keylogging software.

Although an increasing number of researchers study post-editing effort and compare it to translation to determine which one is more productive (Almeida and O'Brien 2010; Guerberof 2009), only a few have analysed post-editing effort as applied to audiovisual translation (de Sousa *et al* 2011; Läubli *et al* 2013), and specifically to subtitling. Other investigations linking audiovisual translation with post-editing have mostly focussed on the quality assessment of machine translated or post-edited subtitles (Armstrong *et al* 2006; Melero *et al* 2006; Volk, 2008; Del Pozo *et al* 2013 or Bywood *et al* 2013).

In order to apply MT and post-editing into the current audiovisual translation workflow, some specificities linked to the genre (wildlife documentary films) and audiovisual transfer modes under analysis (voice-over and off-screen dubbing) need to be taken into account. Voice-over is, together with off-screen dubbing, a modality generally used to translate non-fictional genres in Western Europe (Franco *et al* 2010). Among these non-fictional genres, one can find wildlife documentaries, which form the focus of this research. The main characteristics of documentaries are the presence of both a narrator with a generally planned discourse and experts who tend to use a more spontaneous language (Matamala 2009). Narrators are usually off-screen and dubbed in the target language version, meaning the original narrator cannot be heard and is substituted by a translating voice, whilst on-screen speakers are voiced-over, meaning the translating voice is heard on top of the original, whose sound is lowered down. In both modalities there are synchronisation requirements: translations must take into account the movements and actions on screen (action and kinetic synchronies), and the length of the utterance (isochrony) (Orero 2006). As far as working conditions are concerned, translators sometimes work without a script or with a script riddled with errors due to the possible lack of post-production scripts (Franco *et al*, 2010). All these features may be additional challenges when implementing MT in this specific field, as pointed out in a preliminary study by Ortiz-Boix (forthcoming), which suggested pre-editing, as a necessary step for a more successful implementation of MT. Pre-editing (Pym 1990) is understood as the revision of the format and content of a text before machine translating it. This allows for a higher quality MT output.

3. The experiment: methodological aspects

As stated above, the aim of this experiment was to compare the effort involved in translating and post-editing wildlife documentaries. Following the theoretical approach in section 2, effort was measured in terms of temporal (seconds spent to perform the task), technical (keyboard and mouse usage) and cognitive features (pauses). It was therefore decided that data would be gathered using keylogging software.

3.1. Participants

12 Master students specialising in audiovisual translation participated in this study. They had all taken a specific course on voice-over, in which they were taught to translate documentaries. Tests were carried out in June, when all participants had successfully finished their courses and were working on their MA thesis. Half of the participants were males and the other half were females, ages ranged between 22 and 27 years old, and all of them had completed a BA in Translation and Interpreting. They had minimal or no previous experience as professional audiovisual translators and no experience as post-editors. All participants had Spanish as their first language and were highly proficient in English language.

3.2. Materials

Two excerpts of the 7-minute wildlife documentary *Must Watch: A Lioness Adopts a Baby Antelope* were used. They are available on Youtube as an independent documentary (https://www.youtube.com/watch?v=mZw-1BfHFKM) although it is part of the episode *Odd Couples* from the series *Unlikely Animal Friends* by National Geographic (2009). Both excerpts are comparable in terms of length and content, as shown in Table 1.

		First Excerpt	Second Excerpt
Duration		1:41 minutes (101 seconds)	1:52 minutes (112 seconds)
Words	Total	283	287
	Narrator	50	58
	Experts	222	229
Interventions	Total	8	9
	Narrator	3	3
	Experts	5	6

Table 1. Comparison of excerpts

Both excerpts were machine translated from English into Spanish by *Google Translate* as, according to previous research by Ortiz-Boix (forthcoming), this is the best free online MT engine to translate wildlife documentary films in this language pair. Automatic measures were calculated with the translations and the post-editings produced by the participants (see Table 2 in 5.3.): BLEUs (Papineni 2002), h-BLEU[1]s (Snover *et al* 2006:224), TERs (Snover *et al* 2006) and h-TER[2]s (Snover *et al* 2006:224).

3.3. Data gathering tools

Inputlog (Leijten *et al* 2013), a research tool for logging and analysing writing processes developed at the University of Antwerp, was used to record the data. The following measures were obtained: total time, time spent while performing the task and while searching, keylogging, number of mouse movements and clicks, pause thresholds, type of visited internet webpages and type of used software. Although other post-editing tools were considered, they were discarded because they did not integrate audiovisuals (Ortiz-Boix, *forthcoming*). Inputlog was prioritised over other keylogging software because it allowed for a better simulation of the current workflow of audiovisual translators. It also means that audiovisual materials could be watched without interfering with the tool.

3.4. Test development

Participants volunteered to take part in the experiment, which was carried out in a lab environment simulating real-life working conditions. They were instructed about the nature of the experiment and signed informed consent forms, following the procedures approved by the Ethical Committee at Universitat Autònoma de Barcelona (UAB). They were instructed that the experiment would develop as follows: they would have to translate an excerpt of a wildlife documentary, and post-edit the machine translated output of another excerpt. They were required to use a Microsoft Word template for both tasks, as this was the software used in the MA course they had all taken, but they were free to use any resources available to them online (search engines, video software, etc.). The specific instructions that were given to them were to translate or post-edit, being aware that they had to produce a final document ready to be recorded at a sound studio. They were required to include timecodes in (not out), and they were provided with pre-established timecodes which they could modify if necessary. In the specific case of post-editing, they were instructed to post-edit only when there was a semantic or grammatical error, when some information was omitted or added, and when there were spelling and punctuation mistakes. They were told not to post-edit merely stylistic problems but were asked to rephrase the sentences if, despite being correct, they did not meet the standard conventions of voice-over and off-screen dubbing (this refers to synchronisation features and presentation layouts). After finishing the tasks, they were given a questionnaire on subjective data, the analysis for which is beyond the scope of this paper. Participants were randomly assigned to four different groups in which the two conditions (post-editing/translation) and excerpts (1 and 2) were randomised to avoid any bias regarding the order of presentation.

3.5. Data and methods

20 valid Inputlog files were collected due to technical problems with four files. Data was obtained from the General Analysis Documents file and exported into Microsoft Excel files. They were analysed using the statistical system R-3.1.2, developed at Bell Laboratories by John Chambers and colleagues.

The following data was obtained for all excerpts and tasks:

a) Analysis of temporal effort: average time spent translating and post-editing, average time spent while working on the Word document, on search engines and using video software.
b) Analysis of technical effort: average number of keyboard and mouse usage, average number of mouse movements and scrolls, average number of mouse clicks and average number of keystrokes. Average number of mouse movements and scrolls, mouse clicks, and keystrokes while working on the Word document, on search engines and on video software were also analysed.
c) Analysis of cognitive effort: average number of pauses and average number of pauses while working on the Word document. To

determine PWR and APR, the number of words of each final document and the average time per pause were also assessed.

An ANOVA variance test was used to determine the significance of the results. According to the test, the null-hypothesis can be rejected when the probability value (p-value) is equal or lower than 0.05 (p<0.05). The general null-hypothesis of this research states that "there is a significant difference between post-editing effort and translating effort when working with wildlife documentary films scripts."

4. Results

The global analysis indicates that the post-editing effort is significantly lower than the translating effort in the case of technical effort (F=4.417, p=0.050) and cognitive effort (F=5.979, p=0.025). However, temporal effort is not (F=1.297; p=0.270). This may be due to the time one participant spent post-editing, as he spent nearly double the time the others did. When this participant is not taken into account, the post-editing temporal effort is also lower than the translation temporal effort (F=6.756, p=0.019). Although these results validate our hypothesis, when data from the two different excerpts are analysed in more detail, it can be observed that the difference between post-editing effort and translation effort is not always significant. In the following subsections, and according to the three types of effort identified above, an in-depth analysis is presented.

4.1. Temporal effort

The analysis of temporal effort indicates that, in the first excerpt, participants spent less time post-editing than translating (see Figure 1): the average time spent translating was 2301.833 seconds (38.36 minutes) and 1853.8 seconds (30.9 minutes) for post-editing. The difference between both tasks being 448.033 seconds (7.47 minutes). ANOVA significance test shows that the temporal effort is significantly lower when post-editing (F=12.940; p=0.006), confirming the results of the general analysis.

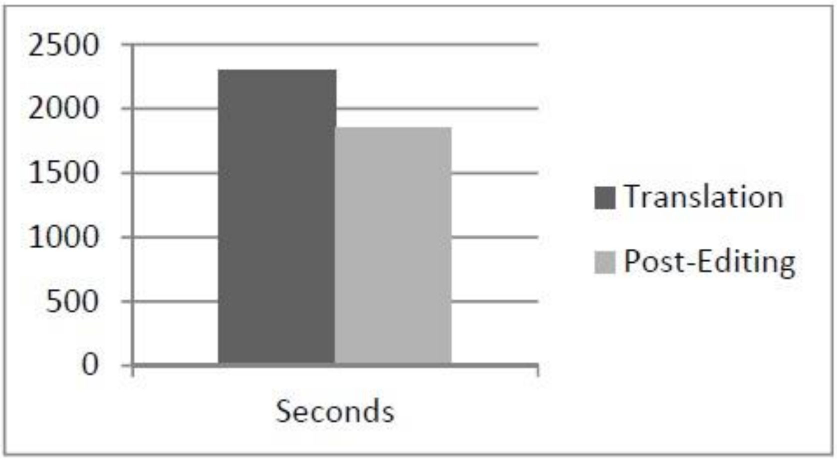


Figure 1. Comparison of Temporal Effort. Excerpt 1.

If the timings are explored in more detail, it can be observed (see Figure 2) that, from all the time dedicated to the performance of the translation, participants spent, in excerpt 1, an average of 1556.1438 seconds (25.94 minutes) on the document (67.605% of the time), 477.0633 seconds (7.95 minutes) on search engines (20.725% of the time) and 152.3562 seconds (2.54 minutes) using the video software (6.619% of the time). When post-editing, the difference between the time performing the task on the document (1137.7662 seconds (18.96 minutes), 61.375% of the time) and on the Internet (378.4386 seconds (6.31 minutes), 20.414% of the time) is smaller. Furthermore, post-editors spent more time using video software (165.263 seconds (2.75 minutes), 8.915% of the time). According to the results, there is evidence leading to the belief that post-editors and translators devote approximately the same time to research (F=1.345; p=0.276) and to the video (F=0.034; p=0.612). However, the time spent on each task within the document is significantly different (F=9.918; p=0.012).

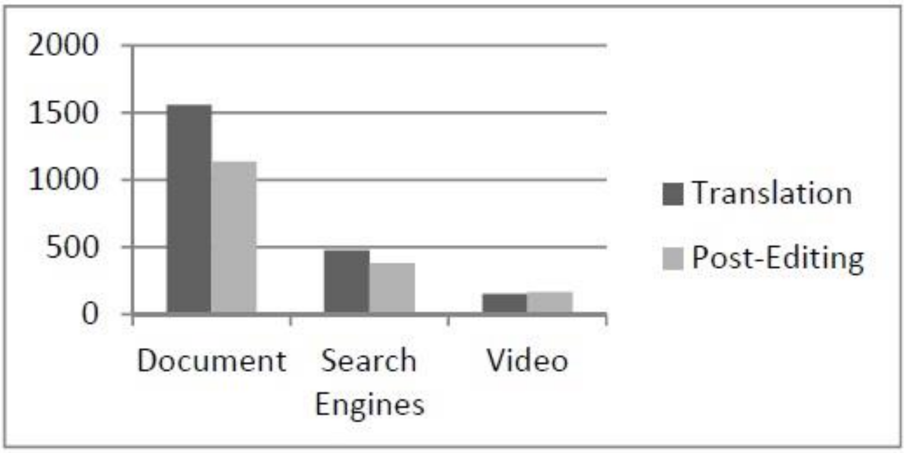


Figure 2. Division of Temporal Effort. Excerpt 1.

In the second excerpt, however, the results of the general analysis are not ratified. In this case, the differences between both tasks are minimal (see Figure 3) and the tendency of greater temporal effort when translating does not continue. The average time for translating is 2054.4 seconds (34.24 minutes) and, for post-editing, 2075.25 (34.59 minutes). This means that it took 20.85 more seconds to post-edit this excerpt. Such a change of tendency, as indicated above, is due to the amount of time one of the participants spent post-editing the excerpt. If this participant is considered an outlier and his data is not taken into account for the analysis, the differences are more similar to those of the first excerpt (see Figure 4): 2,054.4 seconds translating (34.24 minutes) and 1,674.6667 seconds post-editing (27.91 minutes), reversing the difference to 379.7333 seconds in favour of post-editing. In this case, ANOVA significance test (F= 0.002; p=0.965) shows that the difference between post-editing and translation in terms of time is not significant. The difference is closer to be significant when the participant who doubled the time is not included in the data (F= 1.265; p=0.304). As this participant’s behaviour differed considerably from the others, this participant’s results were excluded in the analysis of all the other parameters, which are presented below.

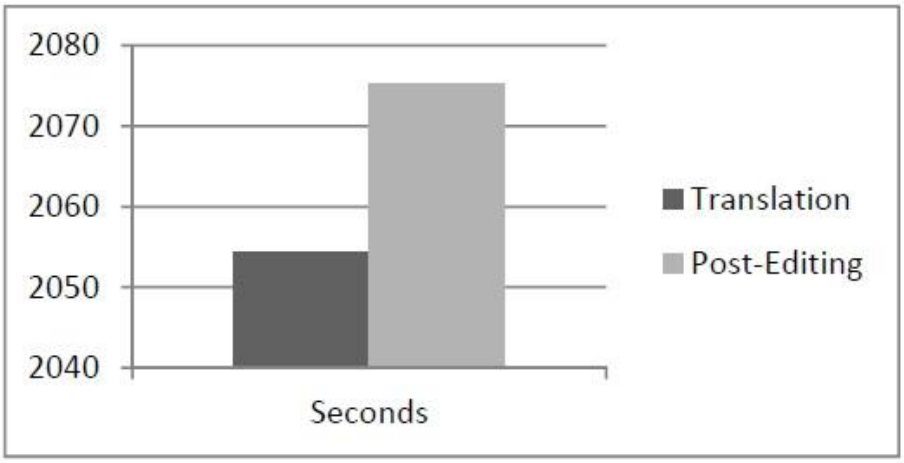


Figure 3. Comparison of Temporal Effort. Excerpt 2.

When the temporal effort for the second excerpt is divided into time spent performing the task within the document, on the search engines or on the audiovisual display, the results are slightly different from the ones obtained in excerpt 1 (see Figure 4). Post-editors spent more time working on the document (1357.577 seconds (22.63 minutes), 81.066% of the time) than translators (1222.78696 seconds (20.38 minutes), 59.520% of the time). Post-editors, however, spent less time on the Internet and using the video software (118.9193 seconds (1.98 minutes), 7.101% of the time, and 122.8303 seconds (2.05 minutes), 7.335% of the time, respectively). Translators spent 328.2352 seconds (5.47

minutes, 15.977% of the time) on search engines and 280.6964 seconds (4.68 minutes, 13.663% of the time) on the audiovisual display. The ANOVA significance test shows that there is no significant difference between translation and post-editing in either the Word document (F= 0.355; p=0.573), the search process (F= 3.480; p=0.111) or when working with the audiovisuals (F= 0.562; p=0.482).

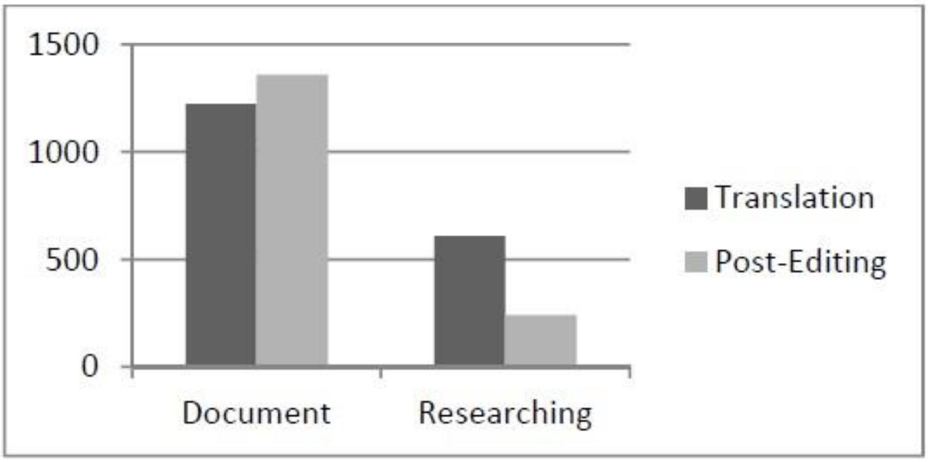


Figure 4. Division of Temporal Effort. Excerpt 2

To sum up, although the general analysis indicates that the post-editing temporal effort is lower than the translation temporal effort, a separate analysis of the two excerpts shows inconsistencies. While in the first excerpt the temporal effort is greater in translation than in post-editing, in the second excerpt there are no significant differences between post-editing and translating in terms of temporal effort. In both, no difference can be seen when considering the time spent when performing the task on the document. However, there is also no significant difference in any of the excerpts when considering the time spent both researching and working with the video.

4.2. Technical effort

The analysis shows that technical effort is higher when translating in both excerpts (see Figures 5 and 6). Translators used the keyboard and the mouse an average of 4079.167 times for the first excerpt and 3972.4 for the second, whilst post-editors used them an average of 2733.8 times for the first excerpt and 2679.333, for the second.

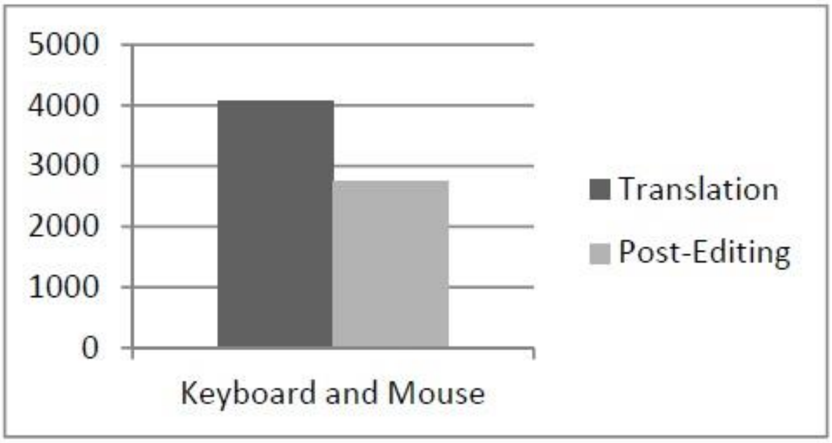


Figure 5. Comparison of Technical Effort. Excerpt 1

In the case of the first excerpt, the difference between the use of technical features when translating and post-editing is of 1345.367 keystrokes and mouse movements and clicks (see Figure 5). For the second excerpt, the difference is a little bit lower (see Figure 6): 1293.067.

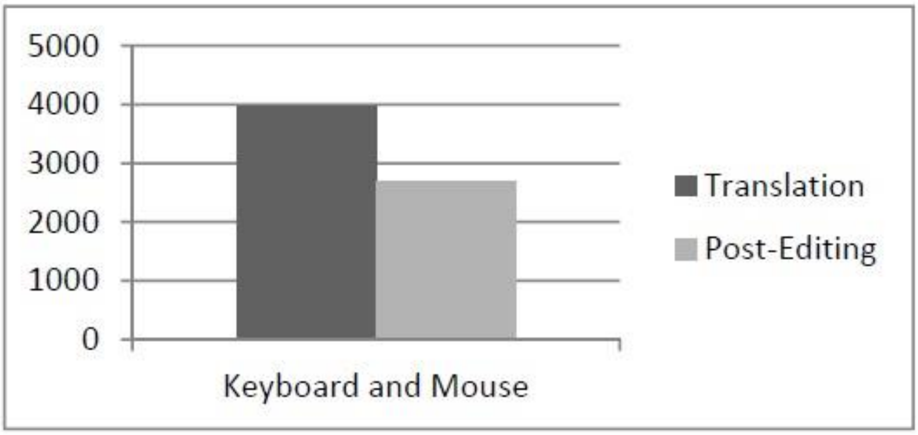


Figure 6. Comparison of Technical Effort. Excerpt 2

According to the results there is evidence to suggest that technical effort is higher when translating than when post-editing. However, the difference is only statistically significant in the first excerpt (F=6.365, p= 0.033; excerpt 2: F=3.529, p=0.109). When technical effort is divided into keyboard strokes and mouse usage, these results show that the difference between post-editing and translating technical efforts is due to keyboard use (F=9.943, p=0.012). While the participants who translated the first excerpt used the keyboard an average of 3183 times and the mouse 896.167 times, the ones who post-edited the same excerpt only used the keyboard 1719 times but moved or clicked the mouse more: 1014.8 times (see Figure 7).

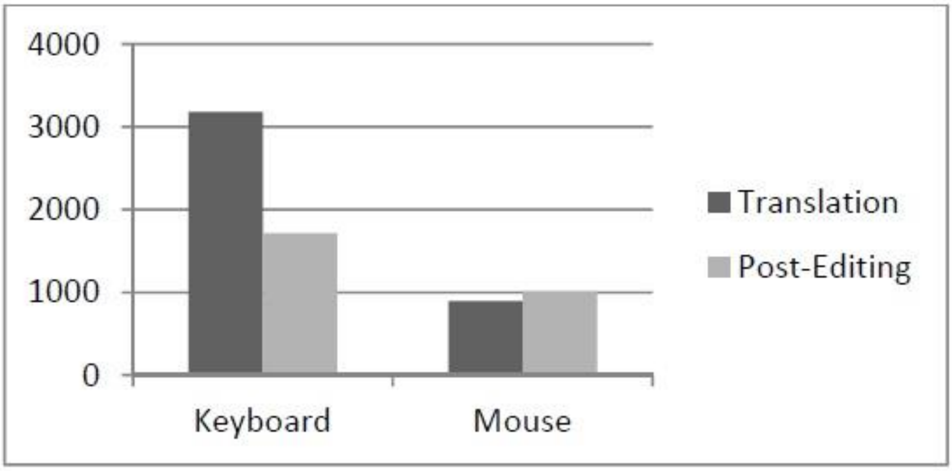


Figure 7. Division of Technical Effort 1. Excerpt 1

The tendency to use the mouse more in post-editing is not followed in the second excerpt (Figure 8). Instead, the participants who translated the second excerpt did so. Translators used the keyboard 3029.2 times and the mouse 943.2 times on average; post-editors made an average of 1974.334 keystrokes and 705 mouse clicks or movements (see Figure 8). Despite the translators making 1,000 keystrokes more than the post-editors, the difference in this case is not significant (F= 4.644, p=0.075).

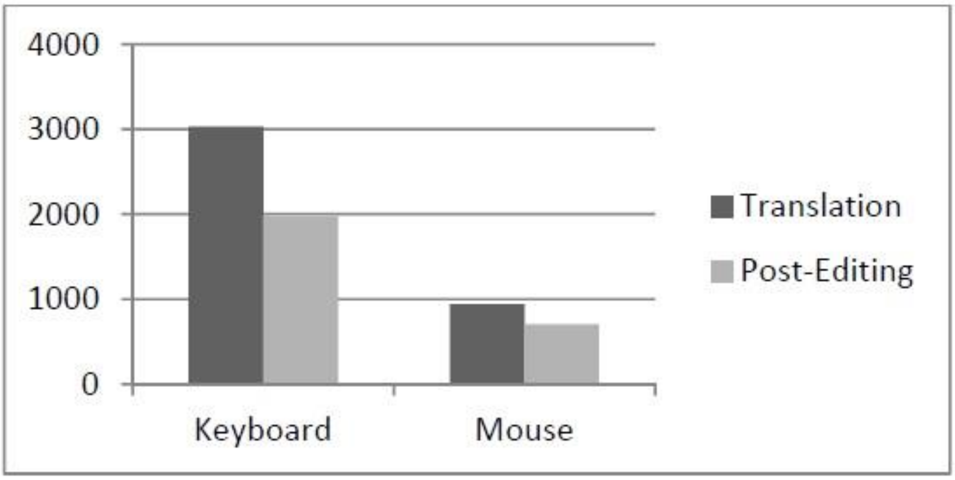


Figure 8. Division of Technical Effort 1. Excerpt 2

When analysing the technical effort distribution in the main document, the search engine and the audiovisual display, one can observe that 79.779% of the technical effort (3254.333 keystrokes and mouse movements and clicks) made by the translators of the first excerpt is concentrated on the main document, 17.802% (726.167 keystrokes and mouse movements and clicks) on search engines and only 2.419% of the effort (98.667 keystrokes and mouse movements and clicks) while using the video software. The post-editors who dealt with the same excerpt dedicated almost the same effort to the audiovisual display (3.382%, 92.4 keystrokes and mouse movements and clicks). Their effort on the main document, 4.679 points lower than the translators' (2051.6 keystrokes and mouse movements and clicks), affected the technical effort while searching on the Internet, which reached 21.517% (587.8 keystrokes and mouse movements and clicks). According to these results, it can be stated that a great majority of the technical effort is concentrated in the main document regardless of the task (see Figure 9).

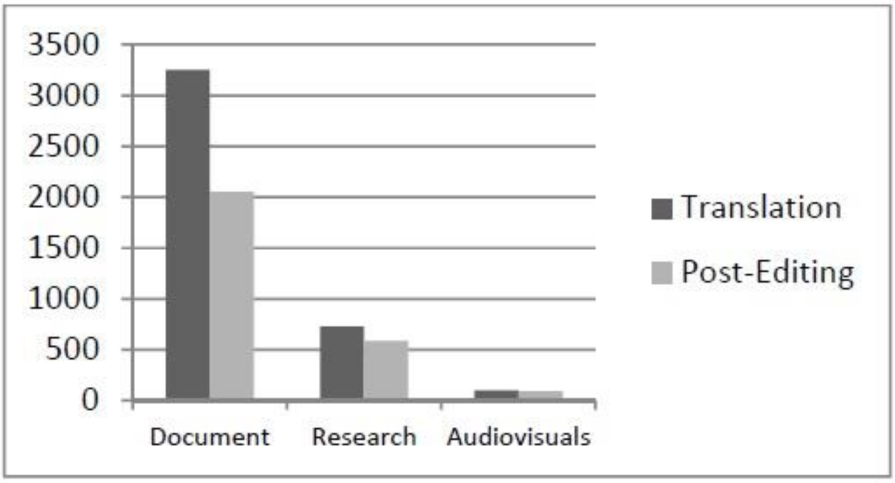


Figure 9. Division of Technical Effort 2. Excerpt 1

The results of the second excerpt follow a similar pattern; technical effort is more concentrated in the document and therefore less technical effort is required where research and audiovisual effort is concerned (see Figure 10): when translating, 81.432% of the technical effort (2420.333 keystrokes and mouse movements and clicks) is concentrated in the main document, while 15.935% (214.667 keystrokes and mouse movements and clicks) is dedicated to the search engines and 2.633% (44.333 keystrokes and mouse movements and clicks) to the audiovisual display. In the case of post-editing, 90.333% of the effort (2234.8 keystrokes and mouse movements and clicks) is made on the document, 8.012% (363 keystrokes and mouse movements and clicks) on the Internet and 1.655% (104.6 keystrokes and mouse movements and clicks) while using the video software.

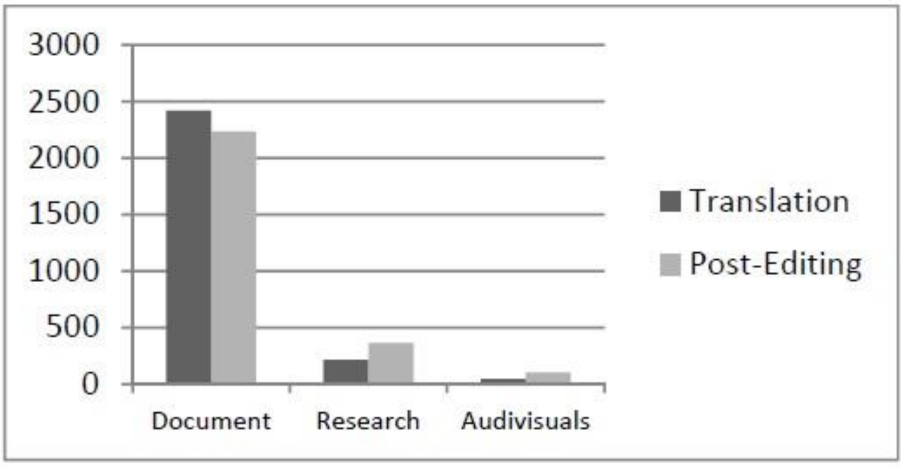


Figure 10. Division of Technical Effort 2. Excerpt 2

Apart from showing that technical effort is basically focused on the main document, the in-depth analysis also shows that when translating and post-editing, the use of the keyboard or the mouse varies: keyboard usage is more intensive when working on the document, while it is almost non-existent when working with the video. When doing online searches, the difference between using the keyboard or the mouse is minimal.

When working within the document, the participants who translated the first excerpt (see Figure 11) used the keyboard an average of 2819.334 times (86.633%) and the mouse, 435 times (13.367%). Translators made an average of 355.833 keystrokes (49.002%) and 370.333 mouse movements and clicks (50.998%) while searching on the Internet; and 78.33 keystrokes (7.939%) and 90.833 mouse clicks and movements (92.061%) while using the video software. The ones who post-edited the same excerpt (see Figure 11) made fewer keystrokes (1419 keystrokes, 69.166%) and used the mouse more extensively (632.6 mouse movements and clicks, 30.834%) while working within the document. In the case of using the search engines and the video software, the difference compared with the results of the translators is minimal. They made an average of 294.6 keystrokes (50.119%) and 293.2 mouse movements and clicks (49.881%), and an average of 3.4 keystrokes (3.679%) and 90.833 mouse clicks and movements (92.061%), respectively.

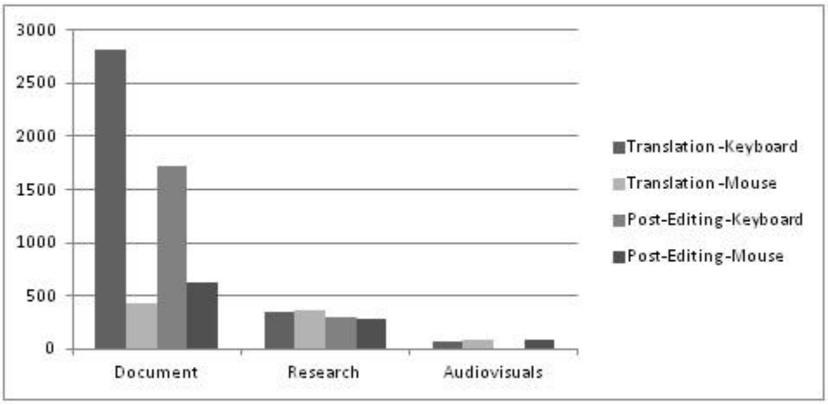


Figure 11. Division of Technical Effort 3. Excerpt 1

Regarding the second excerpt (see Figure 12), the results indicate that the trend continues in the case of working within the document and the video software, but the difference between post-editing and translating with regards to technical efforts while searching on the Internet is a bit higher. On the one hand, the translators used the keyboard an average of 2665.8 times (82.410%) and the mouse 569 times (17.590%), when working within the document. In the case of using search engines, they did 361.2 keystrokes (57.062%) and 271.8 mouse movements and clicks (42.938%). Regarding the technical effort while using the audiovisual display, they used the keyboard an average of 2.2 times (2.103%) and the mouse, 102.4 (97.897%). On the other hand, post-editors made 1,855.333 keystrokes (76.656%) and 565 mouse movements and clicks (23.344%) on the document; and used the keyboard 188.667 times (55.279%) and the mouse 96 times (44.721%) on search engines. In the case of the video software, post-editors used the keyboard an average of 0.334 times (0.752%) and the mouse 44 times (99.248%).

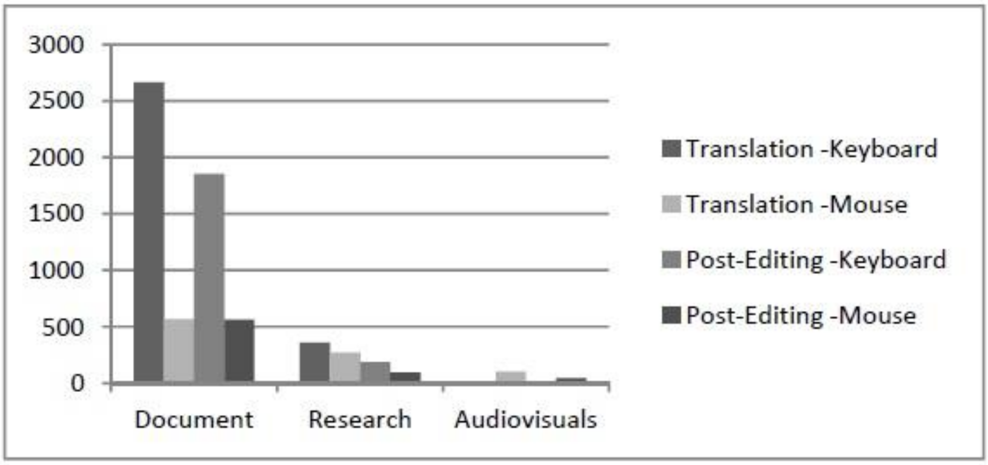


Figure 12. Division of Technical Effort 3. Excerpt 2

To summarise, as in the temporal effort, only the first excerpt follows the trend set by the general analysis, which includes both excerpts. The results show that the improvement of the technical effort is due to the decrease in keyboard usage, which is significantly lower only for the first excerpt. Most of the technical effort is concentrated in the main document, where keyboard usage is more intensive.

4.3. Cognitive Effort

Cognitive effort was assessed using the Lacruz *et al* (2014a) proposal, which states that the higher the difference between APR and PWR, the more cognitive effort is involved. In order to calculate the APR and the PWR for each task and excerpt, two measures gathered by Inputlog were used: total number of pauses and number of pauses while working on the document.

The results obtained for the first excerpt (see Figure 13) showed that the average APR is 0.191301 in the case of translation and 0.244064 for post-editing. The PWR of the same excerpt is 2.947685 for translation and 1.827491 for post-editing. As discussed in section 2, the lower the APR and the higher the PWR, the more cognitive effort is required during the task. Thus, the bigger the difference between APR and PWR, the greater the cognitive effort. The difference between APR and PWR, aka cognitive effort, is significantly higher when translating[3] (total: 2.756384; only document: 2.123383) than when post-editing (total: 1.583427; only document: 1.134013) if the total number of pauses are taken into account (F=11.959; p=0.007) or if only the pauses within the document are considered (F=11.332, p=0.008).

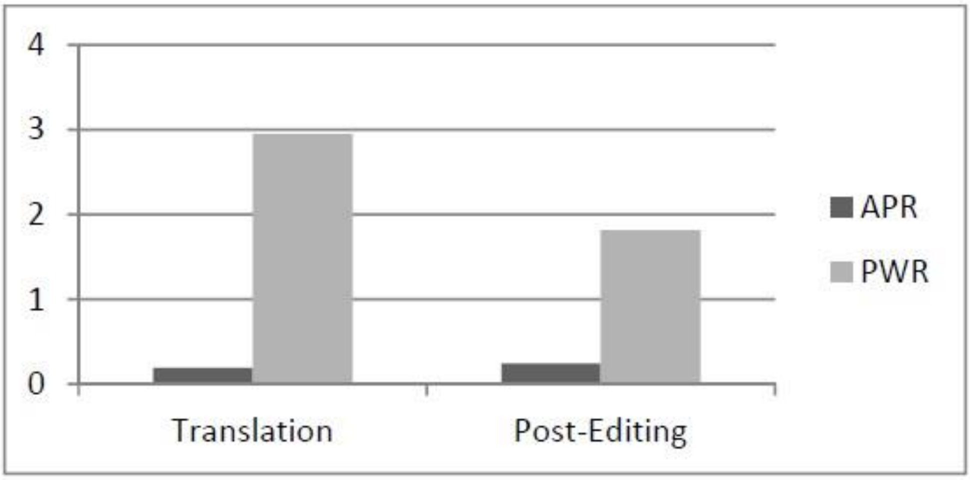


Figure 13. Comparison of Cognitive Effort. Excerpt 1

In the case of the second excerpt (see Figure 14), however, the difference between the translation cognitive effort (total: 1.261884; only document: 1.891389) and the post-editing cognitive effort (total: 1.920086; only document: 2.310353) is not significant even when the total number of pauses are taken into account (F=2.712, p=0.151), or when only the pauses while working within the document are chosen (F=4.155, p=0.088).

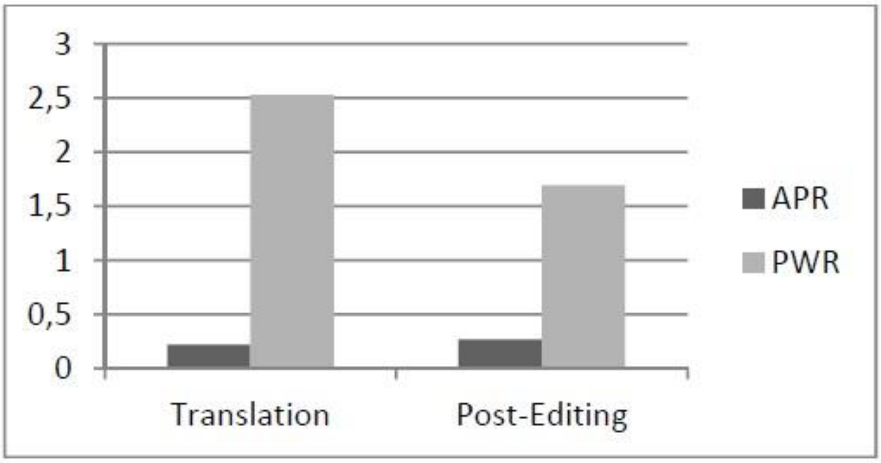


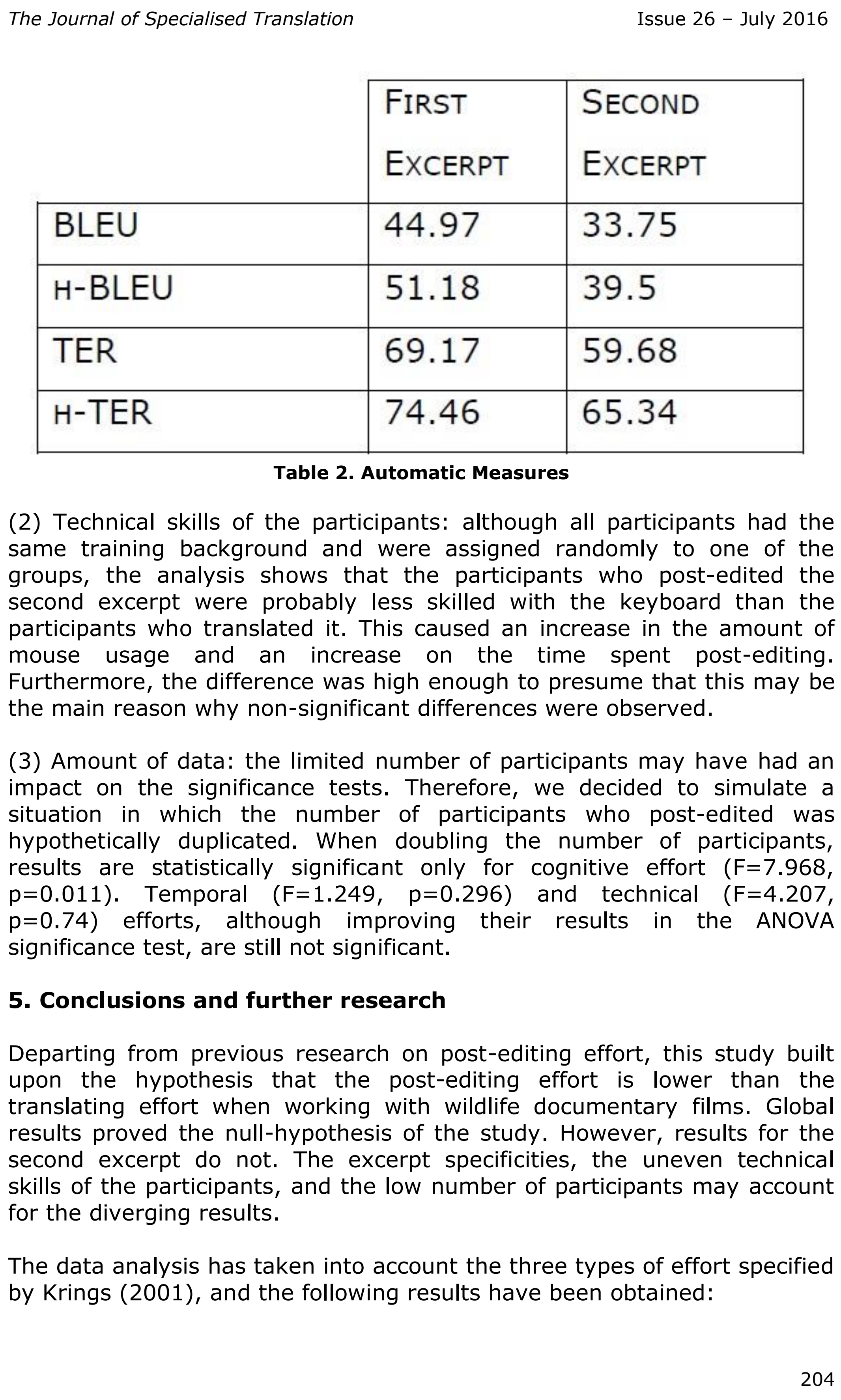
Figure 14. Comparison of Cognitive Effort. Excerpt 2

To sum up, the translation cognitive effort is only significant in the case of the first excerpt. However, although the results of the second excerpt are not significant, the translation cognitive effort is also higher.

4.4. Discussion of results

The results generally confirm the hypothesis that the post-editing effort is lower than the translation effort. Both the general analysis and the analysis of the first excerpt validate the hypothesis, as the temporal, the technical and the cognitive efforts are significantly lower where post-editing is concerned. Nevertheless, the analysis of the second excerpt presents non-significant results. This was unexpected since a previous analysis was carried out to find two comparable excerpts. However, the non-significant results for the second excerpt might be due to three factors:

(1) Features of chosen documentary: although comparable in terms of number of words and interventions, the excerpts were not terminologically and syntactically identical. Furthermore, the MT of the second excerpt was worst, as indicated by the BLEU and TER scores presented (see Table 2).



	First Excerpt	Second Excerpt
BLEU	44.97	33.75
h-BLEU	51.18	39.5
TER	69.17	59.68
h-TER	74.46	65.34

Table 2. Automatic Measures

(2) Technical skills of the participants: although all participants had the same training background and were assigned randomly to one of the groups, the analysis shows that the participants who post-edited the second excerpt were probably less skilled with the keyboard than the participants who translated it. This caused an increase in the amount of mouse usage and an increase on the time spent post-editing. Furthermore, the difference was high enough to presume that this may be the main reason why non-significant differences were observed.

(3) Amount of data: the limited number of participants may have had an impact on the significance tests. Therefore, we decided to simulate a situation in which the number of participants who post-edited was hypothetically duplicated. When doubling the number of participants, results are statistically significant only for cognitive effort (F=7.968, p=0.011). Temporal (F=1.249, p=0.296) and technical (F=4.207, p=0.74) efforts, although improving their results in the ANOVA significance test, are still not significant.

5. Conclusions and further research

Departing from previous research on post-editing effort, this study built upon the hypothesis that the post-editing effort is lower than the translating effort when working with wildlife documentary films. Global results proved the null-hypothesis of the study. However, results for the second excerpt do not. The excerpt specificities, the uneven technical skills of the participants, and the low number of participants may account for the diverging results.

The data analysis has taken into account the three types of effort specified by Krings (2001), and the following results have been obtained:

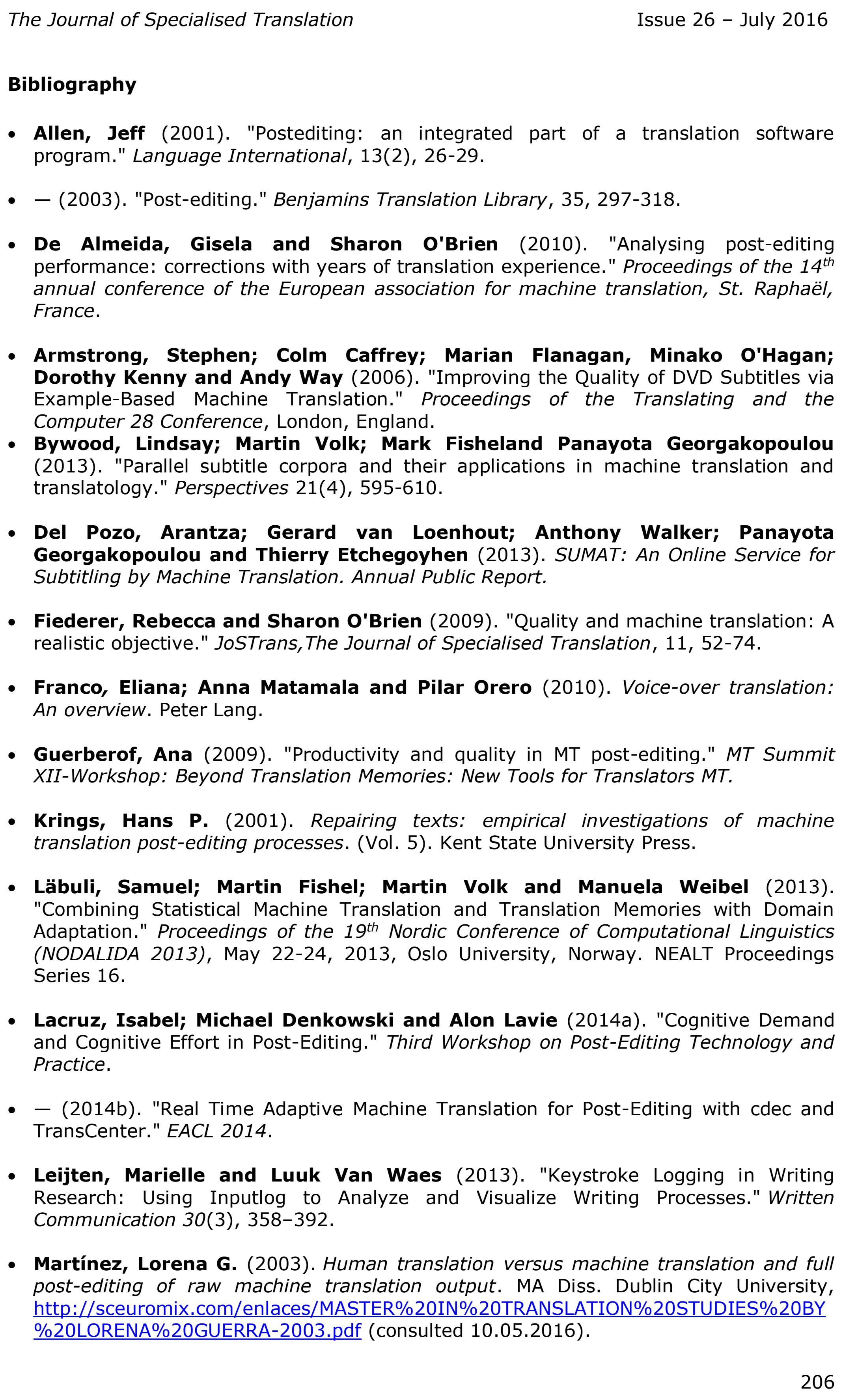
(1) Temporal effort: the global analysis shows that post-editing is faster. However, results are only statistically significant in the first excerpt.

(2) Technical effort: post-editing requires globally less keyboard and mouse usage. Again, the differences are statistically different in the first excerpt but not in the second one.

(3) Cognitive effort: post-editing has been proven to be less cognitively demanding although results are not statistically significant in the second excerpt.

Our data also suggests that the effort is concentrated in the main document and it is precisely there where the effort is reduced. In fact, the effort devoted to the search engines or to the audiovisual display does not vary significantly from one task to the other.

In conclusion, the results seem to indicate that it may be possible to use MT followed by post-editing in specific audiovisual genres such as wildlife documentaries which are voiced-over. However, further research should be carried out to confirm the trends shown in this study, which is limited in scope because it only focuses on one language pair (English into Spanish) and has included a small number of participants. Future research could encompass other types of text and include additional language pairs, with their own specificities. It could also take into account other relevant elements such as the subjective opinions and perceived effort of participants. Other aspects worth researching would be the output quality and audience acceptance of post-edited content in comparison with translated products, along with investigations carried out in other translation modalities (Fiederer *et al* 2009). It would also be highly relevant to measure the professional performance efforts of audiovisual translators. All in all, there are many aspects to be researched but this article has hopefully been a first step towards future studies on the implementation of translation technologies in the field of audiovisual translation and media accessibility, an area that is still under-researched especially when oral modalities such as voice-over, dubbing or even audio description are concerned.



Bibliography

- **Allen, Jeff** (2001). "Postediting: an integrated part of a translation software program." *Language International*, 13(2), 26-29.

- — (2003). "Post-editing." *Benjamins Translation Library*, 35, 297-318.

- **De Almeida, Gisela and Sharon O'Brien** (2010). "Analysing post-editing performance: corrections with years of translation experience." *Proceedings of the 14th annual conference of the European association for machine translation, St. Raphaël, France*.

- **Armstrong, Stephen; Colm Caffrey; Marian Flanagan, Minako O'Hagan; Dorothy Kenny and Andy Way** (2006). "Improving the Quality of DVD Subtitles via Example-Based Machine Translation." *Proceedings of the Translating and the Computer 28 Conference*, London, England.
- **Bywood, Lindsay; Martin Volk; Mark Fisheland Panayota Georgakopoulou** (2013). "Parallel subtitle corpora and their applications in machine translation and translatology." *Perspectives* 21(4), 595-610.

- **Del Pozo, Arantza; Gerard van Loenhout; Anthony Walker; Panayota Georgakopoulou and Thierry Etchegoyhen** (2013). *SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report.*

- **Fiederer, Rebecca and Sharon O'Brien** (2009). "Quality and machine translation: A realistic objective." *JoSTrans,The Journal of Specialised Translation*, 11, 52-74.

- **Franco*, *Eliana; Anna Matamala and Pilar Orero** (2010). *Voice-over translation: An overview*. Peter Lang.

- **Guerberof, Ana** (2009). "Productivity and quality in MT post-editing." *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT.*

- **Krings, Hans P.** (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*. (Vol. 5). Kent State University Press.

- **Läbuli, Samuel; Martin Fishel; Martin Volk and Manuela Weibel** (2013). "Combining Statistical Machine Translation and Translation Memories with Domain Adaptation." *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, May 22-24, 2013, Oslo University, Norway. NEALT Proceedings Series 16.

- **Lacruz, Isabel; Michael Denkowski and Alon Lavie** (2014a). "Cognitive Demand and Cognitive Effort in Post-Editing." *Third Workshop on Post-Editing Technology and Practice*.

- — (2014b). "Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter." *EACL 2014*.

- **Leijten, Marielle and Luuk Van Waes** (2013). "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes." *Written Communication 30*(3), 358–392.

- **Martínez, Lorena G.** (2003). *Human translation versus machine translation and full post-editing of raw machine translation output*. MA Diss. Dublin City University, http://sceuromix.com/enlaces/MASTER%20IN%20TRANSLATION%20STUDIES%20BY%20LORENA%20GUERRA-2003.pdf (consulted 10.05.2016).

- **Matamala, Anna** (2009). "Main Challenges in the Translation of Documentaries." In Jorge Díaz Cintas (ed.). *New Trends in Audiovisual Translation*. Bristol: Multilingual Matters, 109-120.

- **Matamala, Anna ; Anna Fernández-Torné. and Carla Ortiz-Boix** (2012). "Technology and AD: The TECNACC Project." *Languages and the Media 2012*, Berlin. < http://ddd.uab.cat/pub/presentacions/2012/117159/fernandez_matamala_ortiz_berlin 2012.pdf > (last accessed: 16th May 2016)

- **Melero, Maite; Antoni Oliverand Toni Badia** (2006). "Automatic multilingual subtitling in the eTITLE project." *Proc. of the 28th International Conference on Translating and the Computer*, 28 16-17 November 2006 in London. London: ASLIB.

- **O'Brien, Sharon** (2004). "Machine translatability and post-editing effort: How do they relate." *Proc. of the 26th International Conference on Translating and the Computer*, 18-19 November 2004 in London. London: ASLIB.

- — (2005). "Methodologies for measuring the correlations between post-editing effort and machine translatability." *Machine Translation*, 19(1), 37-58.

- — (2006). "Pauses as indicators of cognitive effort in post-editing machine translating output." *Across Languages and Cultures*, 7(1), 1-21.

- — (2010). "Introduction to Post-Editing: Who, What, How and Where to Next." *The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

- — (2011). "Towards predicting post-editing productivity." *Machine Translation*, 25(3), 197-215.

- **Orero, Pilar** (2006). "Synchronisation in Voice-over." *New Spectrum in Translation Studies* 255-264.

- **Ortiz-Boix, Carla** (Forthcoming). "Post-Editing Wildlife Documentaries: Challenges and Possible Solutions."

- **Papineni, Kishore; Salim Roukos; Todd Ward and Wei-Jing Zhu** (2002). "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics (ACL), 311-318.

- **Pym, Peter J.** (1990). "Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system." *Translating and the Computer* 10 (1990), 80-96.

- **Shreve, Gregory M.; Isabel Lacruz and Erik Angelone** (2011). "Sight Translation and Speech Disfluency: Performance Analysis as Window to Cognitive Translation Processes." *Methods and Strategies of Process Research*, Jon Benjamins, 121-146.

- **Snover, Matthew, Bonnie Dorr; Richard Schwartz; Linnea Micciulla and John Makhoul** (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*, 8-12 August 2006 in Cambridge, Massachusetts, USA. Massachusetts: AMTA, 223-231.

- **De Sousa, Sheila C.; Wilker Aziz and Lucia Specia** (2011). "Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles." *RANLP*, 97-103.

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

- **TAUS** (2009). *LSPs in the MT loop: current practices, further requirements*. < https://www.taus.net/think-tank/reports/translate-reports/lsps-in-the-mt-loop-current-practices-future-requirements> (last accessed: 18th May 2016)

- **Veale, Tony and Andy Way** (1997). "Gaijin: A bootstrapping, template-driven approach to example-based MT." *Proceedings of the New Methods in Natural Language Processing (NeMNLP97)*. Sofia, Bulgaria: 239-244.

- **Volk, Martin** (2008). "The Automatic Translation of Film Subtitles. A Machine Translation Success Story?" *Journal for Language Technology and Computational Linguistics*, 23(2), 113-125.

Filmography

- National Geographic (ed.) (2009). "Must Watch: A lioness adopts a baby antelope". *Unlikely Animal Friends*. Episode: "Odd Couples."

Acknowledgements

This article is part of the *ALST project* (reference FFI-2012-31024), directed by Anna Matamala and funded by the Spanish "Ministerio de Economía y Competititvidad," and also of research group TransMedia Catalonia (2014SGR27). This article is also part of the research carried out by Carla Ortiz-Boix under the supervision of Dr Anna Matamala within the PhD in Translation and Intercultural Studies (Department of Translation, Interpreting, and East Asian Studies) at Universitat Autònoma de Barcelona.

The data analysis was carried out during a research stay at Dublin City University. We would also like to thank Dr Sharon O'Brien for welcoming Carla Ortiz-Boix on a research stay at Dublin City University.

Biographies

Carla Ortiz-Boix, BA in Translation and Interpreting (UAB) and MA in Translation, Interpreting and Intercultural Studies (UAB), is a PhD student at the Translation and Interpreting Department at the Universitat Autònoma de Barcelona. Member of the CAIAC research centre and Transmadia Catalona, she was awarded the FI-DGR 2013 pre-doctoral scholarship by AGAUR, an Agency of the Catalan Government.

Email: Carla.Ortiz@uab.cat

Anna Matamala, BA in Translation and Interpreting (UAB) and PhD in Applied Linguistics (UPF). Tenured Lecturer at the Department of Translation and Interpreting at the Universitat Autònoma de Barcelona. Member of the CAIAC research centre and of TransMedia. Audiovisual translator 1996-2007. Anna Matamala has participated in many funded projects on audiovisual translation and accessibility and has published in international journals such as *Meta*, *The Translator*, *Perspectives*, *Babel*, *Linguistica Antverpiensia* and *Jostrans*. She has published a book on interjections and lexicography (2005), and co-authored one on voice – over (with Eliana Franco and Pilar Orero) of a book. She also co-edited three books on audiovisual translation (*Listening to Subtitles*, with Pilar Orero; *Audiovisual Translation in Close-Up*, with Adriana Serban and Jean-Marc Lavaur, and *New Insights in Audiovisual Translation*, with Jorge Díaz-Cintas and Josélia Neves.

Email: Anna.Matamala@uab.cat.

[1] BLEU (Bilingual Evaluation Understudy) and h-BLEU (human targeted Bilingual Evaluation Understudy) are standard automatic measures used to evaluate MT output. The result of these measures arises by comparing MT output with a reference text that can be either its post-editing (BLEU) or a human translation (h-BLEU).

[2] TER (Translation Edit Rate) and h-TER (human targeted Translation Edit Rate) are two other automatic measures used to evaluate MT output. These metrics highlight errors and calculate the edits required in the MT output, in order for the text being edited to resemble a reference text that can be either its post-editing (TER) or a human translation (h-TER).

[3] APR and PWR have been calculated using the total number of pauses and with those pauses being made only in the main document. These two conditions have been selected because the first determines the total cognitive effort and the second specifies cognitive effort within the document where technical effort is the focus.

A3. 3rd Article

Assessing the Quality of Post-edited Wildlife Documentaries

Carla Ortiz-Boix[*] & Anna Matamala[+]

Departament of Translation and Interpretation & East Asian Studies, Universitat Autònoma de Barcelona (UAB)

[*]Carla.Ortiz@uab.cat [+]Anna.Matamala@uab.cat

Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain

Carla Ortiz-Boix, BA in Translation and Interpretation (UAB) and MA in Translation, Interpretation and Intercultural Studies (UAB), is a PhD student at Translation and Interpretation & East Asian Studies at Universitat Autònoma de Barcelona. Member of Transmedia Catalonia and CAIAC Research Center, was awarded the FI-DGR 2013 pre-doctoral scholarship by AGAUR, an Agency of the Catalan Government.

Anna Matamala, BA in Translation and Interpreting (UAB) and PhD in Applied Linguistics (UPF). Tenured lecturer at the Department of Translation and Interpretation & East Asian Studies at Universitat Autònoma de Barcelona. Member of the CAIAC research center and of TransMedia. Audiovisual translator 1996-2007. Anna Matamala has participated in many funded project on audiovisual translation and accessibility and has published in international journals such as Meta, The Translator, Perspectives, Babel, Linguistica Antverpiensia, Jostrans, among others. She has published a book on interjections and lexicography (2005), and is co-autor (with Eliana Franco and Pilar Orero) of a book on voice-over, and co-editor of three books on audiovisual translation (Listening to Subtitles, with Pilar Orero; Audiovisual translation in close-up, with Adriana Serban and Jean-Marc Lavaur, and New Insights in Audiovisual Translation, with Jorge Díaz-Cintas and Josélia Neves).

Assessing the Quality of Post-edited Wildlife Documentaries

This article presents the results of an experiment to assess the quality of post-edited wildlife documentary films to be voiced-over and off-screen dubbed, which was compared to the quality of human translation. The main hypothesis of the article is that there are no significant differences between translated and post-edited texts in terms of quality. Twelve MA students translated and post-edited two excerpts of an English wildlife documentary into Spanish. Then, six professional translators assessed both the translations and post-edited texts by: (1) grading the documents, (2) correcting them using a Multidimensional Quality Metrics-based error classification that takes into account documentary translation specificities, and (3) answering questionnaires on their impressions. Results confirm the main hypothesis by indicating that the quality of post-edited and translated wildlife documentary films is significantly similar.

Screen Translation; Voice-Over; Translation of Documentaries; Machine Translation; Post-Editing; Quality Assessment

1. Introduction

The use of machine translation (MT) followed by post-editing (PE) has been expanding in the translation industry and has been increasingly researched in the last few decades. Several projects investigating the possible inclusion of MT and PE into the process of translating audiovisual products started almost a decade ago. Such projects (e.g. eTITLE or SUMAT) were funded by the European Commission and focused basically on subtitling. The results obtained, especially by the SUMAT project (Etchegoyhen et al, 2014), and the lack of research on the implementation of MT systems in other audiovisual translation modes inspired the ALST project (Matamala et al, 2012) to investigate the possible application of MT and PE into audio description and voice-over, two audiovisual transfer modes which are delivered orally.

The research presented in this article, which is part of the XXX project, focuses exclusively on wildlife documentary films translated by means of voice-over and off-screen dubbing. Voice-over is the revoicing of an audiovisual text in another language in which a translating voice is heard on top of the original voice (Franco et al, 2010:43). It is frequently used in non-fictional audiovisual translation genres, especially when speakers appear on-screen, but also in fictional TV programs in Eastern Europe. Off-screen dubbing, conversely, refers to the audiovisual transfer mode generally used to revoice off-screen narrations in which the original voice is substituted by a target language version (Franco et al, 2010:41).

This article aims to compare the quality of post-edited texts with the quality of human translations and can be considered the follow-up to the investigation presented in Ortiz-Boix & Matamala (forthcoming), where the effort involved in post-editing a wildlife documentary excerpt was compared to the effort involved in translating it. Results showed that the post-editing effort is less than the human translation effort. However, it remains to be seen whether the output quality produced during a post-editing process is similar to the quality produced in a translation process. This article presents the results of an experiment which aimed to validate the hypothesis that there are no significant differences between the quality of post-edited texts and the quality of translations.

The article is structured in six Sections: Section 2 discusses the theoretical approach of the paper. In Section 3, the methodology used in the experiment is explained, as well as the methods used to analyse the data. Sections 4 and 5 discuss the results and, finally, Section 6 presents the conclusions and proposes further research.

2. Quality assessment in translation, machine translation and post-editing

Quality has been a central issue in Translation Studies since the beginning of the discipline and many studies have dealt with it (e. g. Nida, 1964; Carrol, 1966; House, 2006; Koller, 1987; Toury, 1995; Gambier, 1998; Hansen, 2008; Chiaro, 2008). Quality assessment (QA) is product-based and is approached differently depending on the theory of translation that lies behind each QA model (House, 2006). House (2006) divides the different approaches into several categories: (1) anecdotal approaches, which are based on reflections of professional translators who are mainly concerned with the text being faithful to the original (e. g. Savoy 1968); (2) neo-hermeneutic approaches, which consider that the quality of a translation depends on how fully a translator identifies with the original text (e. g. Kupsch-Loseriet 1994); (3) response-oriented approaches, which are communicatively oriented (e. g. Nida 1964, Nida and Taber 1969, Carroll 1966); (4) text- based approaches, which can focus on linguistic (Reiss 1978), literary (Toury 1985) or functional aspects (Reiss and Vermeer 1984) of the translation, and (5) pragmatic linguistic approaches, which analyse the linguistic-situational particularities of source and target texts, compare them and assess their relative match to assess quality (House, 2006). Depending on the approach, various methods have been proposed, such as quantitative and qualitative testing by competent judges, comparing translations against reference models, rating sentences according to pre-established scales of intelligibility and informativeness and gathering respondents' opinions, among others.

Studies on MT have also addressed quality and QA as a main issue (see e.g. Hutchins & Somers, 1992; Krings, 2001; King et al, 2003; Fiederer et al, 2009; Armstrong et al, 2006). While the first studies only used human judges to evaluate the MT output, measures to assess the quality automatically by means of a preprogrammed tool (automatic measures) have been expanding. Such measures compare and correlate translations or post-edited texts with the MT output to set the quality of the machine translated texts. Thus, automatic measures still need human translations: on the one hand, there are measures of precision such as BLEU (Papineni et al, 2001), H-BLEU (Snover et al, 2006), NIST (Snover et al, 2008) or METEOR (Lavie et al, 2009), which compare the MT output with reference translations or post-edited texts. On the other hand, there are editing-distance measures such as TER (Snover et al, 2006), H-TER (Snover et al, 2006), WER (Jiménez Linares, 2008) or PER (Jiménez Linares, 2008), which calculate the number of modifications needed on a MT output so that it resembles a reference translation or post-editing.

Apart from automatic measures, human-based evaluations have been carried out on raw MT output (e. g. O'Brien, 2005; Aziz et al, 2012). The majority of these evaluations focus on the fidelity or accuracy of the MT output (e. g. Arnold et al, 1994; Dabbadie et al, 2002; Roturier, 2006; Fiederer, 2009), its intelligibility or clarity (e. g. Hutchins and Somers,1992; Fiederer et al, 2009), and its style (Hutchins and Somers, 1992; Arnold et al, 1994;Fiederer et al, 2009). Others focus on using post-editing as a measure of assessment (e.g. Popovic et al, 2013) or on classifying the errors produced by the MT engines (e.g. Federman, 2012). Regarding error classification, Lommel et al (2014) designed the Multidimensional Quality Metrics (MQM), based on functional theories of translation, which propose several error issue types dealing with both the micro and the macrostructure of the text. This metrics can be used to assess MT output, but also post-edited texts and translations (Lommel et al, 2014).

Regarding research on the specific topic of our investigation –evaluating the quality of post-editing in comparison to translations– it is rather limited. Plitt et al (2010), Guerberof (2009, 2012), Fiederer et al (2009), Carl et al (2011), García (2011) and De Sutter et al (2012) have compared, to a greater or lesser extent, the quality of post-edited texts and translations.

Plitt et al (2010) assessed the quality at Autodesk, a company whose Localization Service department actively uses MT and measures the usefulness of MT, which can be used as a translation productivity tool or for gisting. Autodesk's translation QA team reviewed part of the work of 10 out of 12 participants, who translated and post-edited randomly selected samples of translated and post-edited texts from English into French, Italian, German and Spanish, and rated them in two levels: "average" or "good", depending on whether they would publish the texts as they read them. The results presented in Plitt's article showed that translations contained a higher number of mistakes than post-edited sentences in all four languages.

Guerberof (2009, 2012) used three reviewers to blindly assess human translated and post-edited MT segments, as well as corrected segments previously extracted from a translation memory, on the topic "business intelligence technology" by using the LISA QA model. These three reviewers measured and classified the number of errors in eight categories: mistranslation, accuracy, terminology, language, style, country, consistency, and format. During the analysis, the number of errors per source (human translated segments, post-edited MT segments or corrected segments extracted from a translation memory) was calculated. Guerberof's (2012) results show that the quality produced by translators was significantly higher when they post-edited a segment produced by the MT engine or proposed by the translation memories. It was also observed that while the majority of language, terminology and style errors were found in the segments translated from scratch, the majority of accuracy errors were seen in the corrected segments extracted from translation memories, and mistranslation errors were mainly present in post-edited MT segments.

As part of a wider study, Fiederer et al (2009) assessed the quality in machine-translated texts by evaluating 30 source sentences with three translated and three post-edited versions according to three parameters: clarity, accuracy and style. The sentences were assessed by 11 raters, who ranked the translations and post-edited texts from 1 to 4 (being 1 the lowest mark and 4 the highest). They were also asked to indicate their favorite translated option out of the six proposals for each source sentence. Evaluators scored translated and post-edited texts equally in terms of clarity. However, post-edited texts were evaluated higher with regards to accuracy, and translations were evaluated higher when style was considered. All in all, raters chose the translated sentences as their favorites.

Carl et al (2011) presented a study on the post-editing experience of translators working from English into Danish. It included the QA of three general texts (850 characters all together), evaluated by seven native speakers of Danish. Each rater ranked two human translations and two post-edited texts. Post-edited texts were found to be better than translations, although the difference was not statistically significant.

García (2011) explored post-editing in non-professional contexts in the English-Chinese language pair. In order to do so, one rater assessed the quality of a 500-word text both translated and post-edited by using the Australian National Accreditation Authority for Translators and Interpreters' (NAATI) guidelines. The results presented in the study show that post-edited passages were of higher quality than the translated.

Finally, De Sutter et al (2012) studied the quality of a text half translated half post-edited by 15 translation trainees from English into French. The assessment was done by a single evaluator who rated the segments using a five-point scale. The results show that translations receive higher scores than post-edited texts, although the difference between the translations and post-edited texts was minimal.

Taking into account all this prior work on the topic, a mixed-approach model for QA, based on both text-based and response-oriented models, has been proposed for the current experiment, as described next.

3. Methodology

The aim of the experiment is to assess the quality of post-edited wildlife documentaries compared to the quality of human translations. It is built upon the hypothesis that there is no significant difference between the quality of post-edited and the quality of translated wildlife documentaries to be voiced-over and off-screen dubbed from English into Spanish.

3.1. Participants

The evaluators participating in this study were six lecturers of audiovisual translation MAs in Spanish universities who are experts on voice-over and currently work or have worked as professional voice-over translators. Participants' profiles are comparable, as all of them have a BA in Translation Studies except for one, who has a BA in German Philology. Five of them have either attended PhD courses or have a PhD in Translation. Furthermore, they all work from English into Spanish. Two participants are more experienced than the others: when the experiment was carried out, participants 1 and 5 had worked as audiovisual translators for 16 years and had taught for 11 and 5 respectively. In comparison, participants 2, 4 and 6 had between 5 and 8 years of experience as audiovisual translators and had taught for the last 4 or 5. Participant 3 had worked as audiovisual translator for 10 years and taught for the last 8. The number of raters is limited but in line with previous research in QA of MT output using human judges (e. g. De Sousa et al, 2011) and even higher than existing post-editing experiments (Guerberof, 2009; Guerberof, 2012; García, 2011 and De Sutter et al, 2012).

3.2. Materials

The materials used were 6 translations and 6 post-edited texts of two excerpts of the 7-minute wildlife documentary film *Must Watch: a Lioness Adopts a Baby Antelope* (i.e. a total of 24 documents*)*. The translations and post-edited texts were produced by 12 students of an MA in audiovisual translation who had taken a specific course in voice-over where they were taught how to translate documentaries. The documentary is an excerpt from the episode *Odd Couples* from the National Geographic series *Unlikely Animal Friends* broadcast in 2009 and currently available as an independent video on YouTube[1]. The excerpts are similar in length, number of words and entries (Table 1). A short transcription of one of the excerpts used for the experiment is included in the following lines as an example of the type of text used:

[1] https://www.youtube.com/watch?v=mZw-1BfHFKM (Last accessed: 24th February 2016)

00:00 ***Narrator:*** *For days the calf wandered looking for its herd, while the lioness followed.*

00:06 ***Saba:*** *Of course, every oryx it saw was potentially its mother and potentially food and life. So, it would constantly try to rejoin adult oryxes. Well, Kamunyak would allow it to go a certain distance away but as soon as it started to move off with the oryxes she was then up on the warpath.*

[...]

Both excerpts were machine translated by *Google Translate*, as a pre-analysis by Ortiz-Boix (forthcoming) proved this was the best free online MT engine that can be used to translate wildlife documentary scripts at the time the experiment took place. The pre-analysis compared the output produced by Google Translate to the output of 7 other free online MT engines using automatic quality measures and a human analysis of errors. The MT of the first excerpt was slightly better than the MT of the second, according to automatic metrics such as BLEU[2], h-BLEU[2] (Papieni, 2002), TER[3] and h-TER[3] (Snover et al, 2006) (see also Table 1).

[2] BLEU (Bilingual Evaluation Understudy) and h-BLEU (human targeted Bilingual Evaluation Understudy) are standard automatic measures used to evaluate MT output. The result of these measures arises by comparing MT output with a reference text that can be either its post-editing (BLEU) or a human translation (h-BLEU). The given result is between 0 and 100, where the higher the score is, the better the translation is considered.

[3] TER (Translation Edit Rate) and h-TER (human targeted Translation Edit Rate) are two other automatic measures used to evaluate MT output. These metrics highlight errors and calculate the edits required in the MT output, in order for the text being edited to resemble a reference text that can be either its post- editing (TER) or a human translation (h-TER). The given result is between 0 and 100, where the lower the score is, the better the translation is considered.

Table 1. Comparison of excerpts

3.3. Test development

Participants carried out the experiment from their usual place of work. They were given detailed instructions to assess 24 documents in 20 days without knowing which of them were translations or post-edited texts. The experiment was divided in two parts:

(1) For the first part they were given just one day: they were instructed to read each document and grade it according to their first impression on a 7-point scale (scoring round 1). The order of the documents was randomized. It was expected that this approach would reflect more accurately how the audience would react to the documentary and provide interesting findings on the quality of the text from a target audience perspective.

(2) In the second part, they were asked to review and correct the documents, identify the errors following a specific evaluation matrix (see Section 3.4) and grade the documents after the correction on a 7- point scale (scoring round 2). The approach in this second round was more academic, as reviewers were given a set of specifications that guided them through a more didactic analysis. Afterwards, they were instructed to answer an online questionnaire on their opinion about the document they corrected (questionnaire-based evaluation, see Section 3.4.), and gave each document a final grade between 0 and 10 (scoring round 3). These two last grading rounds intended to assess the overall quality after having completely analysed the text and reflected on every aspect of the translation or post-editing. They also had to guess whether the assessed document was a translation or a post-

editing (post-editing/translation identification task). Evaluators were allowed to complete the assessments at their own pace during the 20 days allotted, as long as the correction of each document was done within the same day.

3.4. Evaluation matrix and questionnaire design

The evaluation matrix used for this study is based on the Multidimensional Quality Metrics (MQM) established by Uszkoreit et al (2013), as MQM is designed to assess both human and machine translations, and allows us to check only the categories that are considered relevant for each specific text type in as much or little detail as needed. MQM also permits the inclusion of other categories dealing with domain specific issues. Although MQM offers over one hundred categories and subcategories of issue types, only five categories and eleven subcategories were selected (see Table 2). The selection was based on previous research on the most common MT engines errors both in general texts (Avramidis et al, 2014) and in wildlife documentary films (Ortiz-Boix, forthcoming), and also in post-edited texts (Guerberof, 2009). Furthermore, as MQM does not include any audiovisual translation domain specific issues, a new category containing four subcategories was included: voice-over/off-screen dubbing specificities (see Table 2). These specificities were: spotting[4], action synchrony, voice-over isochrony (Franco et al, 2010), and inclusion of phonetic transcriptions to facilitate the pronunciation of foreign names by voice talents. Although these elements would not be relevant when analysing machine translation output, they are considered in the current experiment, where the quality evaluation is done not on

[4] Spotting, also called timing or cueing, is the process of defining in and sometimes out time codes of each voice-over or off-screen dubbing unit (Ortiz-Boix, forthcoming).

the machine translation output but on post-edited texts that have to fulfill certain requirements of the audiovisual transfer mode under review. Table 2 summarises the evaluation matrix used. The identification number that will be later used in the analysis for each category is included in the right column.

Table 2. Evaluation matrix: error typology

Evaluators were given an explanatory document with a definition and example of each issue type. Prior to the experiment, a pilot test and specific training were carried out to confirm the categories were appropriate and procedures were correctly understood.

As for the questionnaire, the participants had to report their level of agreement with eight statements on a 7-point Likert scale, where one equated to “completely disagree” through seven equating to “completely agree”. The statements were the following:

(1) Generally speaking, the text was fluent;

(2) Overall, the text was grammatically correct;

(3) Broadly speaking, there were no spelling mistakes;

(4) Generally speaking, the vocabulary was appropriate;

(5) The vocabulary was mostly coherent throughout the text;

(6) In general, the text fulfills the standards of voice-over translation;

(7) Overall, the final result was satisfying;

(8) The text could be sent to a dubbing studio to be voiced-over.

3.5. Data and methods

After all the evaluators performed their tasks, one hundred and forty-four corrected documents were collected along with the corresponding questionnaires. They were analysed using the statistical system R-3.1.2 (https://www.r-project.org/), developed by John Chambers and colleagues at Bell Laboratories. The following data were obtained and analysed:

(1) The grades for each document in the three scoring rounds. For round 1 (after reading the document for the first time) and round 2 (after correcting the documents, before answering the questionnaire), the following scale was used: "completely unsatisfactory", "deficient", "fail", "pass", "good", "very good" and "excellent". For round 3 (after correcting the documents and answering the questionnaire), a more precise scale similar to the ones lecturers apply in their courses' assessment was used: a numerical scale from 0 to 10, being 0 the lowest mark and any mark below 5 equal to a "fail". The data for rounds 1 and 2 are discussed in Subsection 4.1 globally and in Subsection 4.3. separately for each excerpt. The data for round 3 can be found in Section 4.5.

(2) 144 questionnaires (6 x 24 documents) reporting on the participants' opinions after correcting each document. An analysis of the questionnaire replies is provided in Section 4.2. (globally) and 4.4 (separately for each excerpt).

(3) The results of the post-editing/translation identification task. This issue is discussed in Subsection 4.6.

(4) 144 documents with corrections according to the evaluation matrix based on the MQM: 6 corrected documents for each of the 24 documents. The number and type of errors corrected are discussed in Subsection 4.7.

An ANOVA variance test was carried out to validate the hypothesis. Statistical significance is assumed for $p<0.05$, meaning that the difference between the results of the post-editing and translation QA should be higher than 0.05 to be considered significant.

4. Discussion of results: scoring rounds and questionnaire replies

This Section discusses the global results taking into account scoring rounds 1 and 2 and the questionnaires replies (Subsection 1), and it then considers each excerpt separately (Subsection 2). Next, it presents the results of scoring round 3 (Subsection 3). It finally discusses the post-editing/translation identification task (Subsection 4).

4.1. Global analysis

The results of the global analysis, which includes the data from both excerpts, indicate that the differences, in terms of quality, between translation and post-editing of wildlife documentary films are not high. Thus, these results seem to validate the hypothesis of the study, as evaluators consider translations and post-edited texts qualitatively comparable both in the case of the grades given in the scoring rounds (see Figure 1) and in the questionnaire-based evaluation (see Figure 2).

Figure 1. Global evaluation results: scoring rounds

Furthermore, when focusing only on the evaluators' first scoring round, translations are better than post-edited texts, as 62.5% of translations (45 out of 72) have been evaluated from “pass” to “excellent” whilst only 51.38% of the post-edited texts have (37 out of 72). Furthermore, when focusing only in the best-rated outputs (from “good” to “excellent”), translations also get better scores (27 vs. 21). However, the median value for both translations and post-editing in round 1 is the same: “pass”.

As far as round 2 is concerned, the difference between translations and post-edited texts is reduced: 56.95% (41 out of 72) of the translations and 52.78% (38 out of 72) of the post-editings were evaluated from “pass” to “excellent”. When only considering those between “good” and “excellent”, 33.30% of the translations (24) and 29.17% of the post-edited texts (21) are found in this range. Even narrowing the scope to the best outputs, the number of translations that can be included in the ranges between “good” and “excellent” is higher than the number of post-edited texts included in them. However, when descriptive statistics are performed, it can be seen that the median grade for both tasks is “pass”.

Comparing the results of round 1 and 2, results are lower in round 2, which might lead to the conclusion than the more in depth the raters assess, the stricter they are and fewer differences between translation and post-editing are observed. The strictness of the second round might be due to the fact that evaluators had assessed the translated and post-edited documents in a more didactic way and according to a set of specifications. Hence, they could be aware of problems that had been not noticed during the first round, when they evaluated the translated and post-edited texts globally, adopting a more audience-centric perspective. Additionally, although the results of rounds 1 and 2 seem to indicate that translations are better than post-edited texts, the results are not statistically significant (1st round: F=0.000, p=1.000; 2nd round: F=1.000, p=1.000)[5], leading to the conclusion that post-edited texts and translations are significantly similar in terms of quality.

When adopting a different evaluation system, the results are slightly different. The questionnaire-based assessment indicates post-edited texts are better than translations: translations are given lower grades in 4 out of 8 specific evaluation issues – grammar, coherence, correction and adequacy of the text so that it can be sent to a dubbing studio- and the same grade in another item – VO specificities (see Figure 2).

Figure 2. Global evaluation results: questionnaire-based assessment (mean values)

In this case, the evaluators scored the issues from 1 to 7, with 1 being the lowest grade. In those instances where translations got higher grades, mean grades for post-edited texts are no more than 0.1 points lower. The largest difference (0.247 points) can be found for issue type "coherence", where post-edited texts get a better grade. And the smallest difference is for "voice-over specificities", where there is no difference between translations and post-edited texts (4.181). However, in all categories the difference between post-edited texts and translations is again non-significant: fluency (F=0.155, p=0.695), grammar (F=0.004, p=0.948), spelling (F=0.691, p=0.407), vocabulary (F=0.019, p=0.892), coherence (F=0.410, p=0.523), VO specificities (F=0.000, p=1.000), correction (F=1.450, p=0.230), dubbing studio (F=0.581, p=0.447). When the questionnaire-based assessment is correlated

[5] "F=" stands for "F-value", which shows if a group of variables are significant together. "p=" stands for "p-value", which shows the provability of obtaining an equal or similar result to what has been observed in this particular experiment and, hence, its significance.

with the global analysis, results indicate that issues related to terminological coherence, grammar and dubbing studio specificities have more impact on the grades than issues related to spelling, vocabulary or fluency, as the difference between translations and post-edited texts shortens after evaluation round 2.

Results show, therefore, that in general there is no significant difference between the quality of post-edited texts and translations in the analysed excerpts.

4.2. Specific analysis

When the results for each excerpt are analysed separately, some differences appear, even though the results are not statistically significant here either. For excerpt 2, results indicate that the quality of post-editing and translation after evaluations rounds 1 and 2 is almost equal (see Figure 3).

Figure 3. Specific evaluation results: scoring rounds (excerpt 2)

In round 1, translations are slightly better than post-edited texts, as 61.11% of the translations (22) versus 50% of the post-edited texts (18) get a pass grade (from “pass” to “excellent”). However, when focusing on the outputs in the range between “good” and “excellent”, the number of post-edited texts and translations is the same (30.55%, 11 out of 36), and there are more post-edited texts rated as “very good” than translations (6 versus 1). However, the median for both translations and post-edited texts is again the same: “pass”. In round 2, if data are divided into two groups (fail and below/pass and above), the percentage is exactly the same for both post-edited texts and translations: 44.45% (16 out of 36) vs 55.55% (20 out of 36). However, when looking at the distribution in the higher range, one can observe that translations get higher marks than post-edited texts. Again, though, the median grade for both tasks is “pass”, showing no significant differences with round 1. It must be noticed that raters were stricter in round 2 and less differences were found between translations and post-edited texts. Results in both rounds were not statistically significant (round 1: F=0.584, p=0.447; round 2: F=0.004, p=0.748), confirming that post-edited texts and translations in our experiment are quite similar in terms of quality.

For excerpt 1, however, results present wider differences between post-edited texts and translations, as shown in Figure 4.

Figure 4. Specific evaluation results: scoring rounds (excerpt 1)

In round 1, translations seem to be better than post-edited texts, as 63.89% (23 out of 36) get pass grades (from “pass” to “excellent”) compared to 52.77% of post-edited texts (19). The difference is still more striking in the higher marks: 44.4% of translations (16 out of 36) get between “good” and “excellent”, whilst only 27.78% (10 out of 36) post-edited texts are found in this range. However, statistics show that the median score for both conditions is the same: “pass”. In round 2, the difference between translations and post-edited texts evens out, as 58.33% (21 out of 36) and 50% (18) of post-edited texts get pass grades. In the higher marks, the difference is 33.33% translations (12) versus 22.22% post-edited texts (8). The median grade for translation is “pass”, whilst the median for post-editing falls between "pass" and "fail", again showing the tendency of evaluators to be stricter in second rounds. When inferential statistics are performed, results show again no significant differences in both rounds (round 1: F=0.584, p=0.447; second round: F=0.004, p=0.948), confirming the conclusion that post-edited texts and translations are significantly similar in terms of quality.

To summarize, although the results of the global analysis (including both excerpts) indicate that translations receive better marks than post-edited texts, the difference between them is not statistically significant. When the results are divided according to the excerpts, opposing trends are observed: in excerpt 2 post-edited texts receive better grades than translations, whilst in excerpt 1 translations receive higher marks. However, differences are not statistically significant in any of the excerpts under analysis. The results also indicate that the difference between translation and post-editing narrows after each round of evaluation.

As for the questionnaire responses, the biggest differences are found in the second excerpt (see Figure 5), where post-editing received better grades than translation in four of the specific evaluation issues. However, these differences are not high, with coherence being the issue type where the widest difference is to be observed (4.25/4.667).

Figure 5. Specific evaluation results: questionnaire-based assessment (excerpt 2)

Although the grades for post-edited texts are generally higher than for translations, such differences are not statistically significant in any case: fluency (F=0.000, p=1.000), grammar (F=0.254, p=0.616), coherence (F=3.182, p=0.079), correction (F=2.248, p=0.138), and adequacy for the dubbing studio (F=0.506, p=0.479). When translations perform better, differences are again non-significant: spelling (F=0.103, p=0.749), vocabulary (F=0.042, p=0.839) and VO specificities (F=0.100, p=0.753).

The results for the first excerpt, however, again present much narrower differences (see Figure 6).

Figure 6. Specific evaluation results: questionnaire-based assessment (excerpt 1)

Translations are better than post-edited texts in five aspects, although the differences are not significant in any of the cases: fluency (F=0.266, p=0.608), grammar (F=0.267, p=0.607), spelling (F=0.735, p=0.394), vocabulary (F=0.130, p=0.719) and coherence (F=0.608, p=0.438). In other aspects, such as VO specificities (F=0.108, p=0.743), correction (F=0.079, p=0.780) and adequacy for the dubbing studio (F=0.140, p=0.709), post-edited texts are evaluated as better. The difference between the mean grades is not significant in any aspect, which leads to the belief that translation and post-editing are comparable.

To sum up, in all cases, translation and post-editing are significantly similar. Although there are differences between translations and post-edited texts in both excerpts, such differences are minimal and, therefore, the results prove the null-hypothesis of the article is correct.

4.3. Third round of evaluation

After answering the questionnaire, evaluators graded each text for the last time (round 3). In this case, they gave translation and post-edited texts a grade from 0 to 10, being 0 the lowest grade and 10 the highest, as this is the scale lecturers use to evaluate at university. Figure 7 presents the mean results of this evaluation.

Figure 7. Evaluation round 3

As far as the global final grade (including both excerpts) is concerned, mean grades are 5.3505 for post-editing and 5.448 for translation, a difference which is not statistically significant (F=0.000, p=1.000). When each excerpt is analysed independently, results are also better for translations, although the difference is higher for the second excerpt (0.153) than for the first (0.043). Differences are again not statistically significant (F=0.000, p=1.000; F=0.001, p=0.975), which is in line with results in previous scoring rounds. Similarly, the slight differences found between translations and post-editing narrow in each of the evaluation rounds, with the difference between round 1 and 2 the highest.

4.4. Identification of post-edited texts and translations

Evaluators were asked to identify each corrected document as a translation or as a post-editing. The results show that it is easier to assert which ones are translations (see Figure 8), as 42 out of 72 translations (58.33%) were correctly identified and only 14 translations (19.44%) were wrongly identified as post-edited texts. The remaining 16 translations (22.22%) were not identified by the evaluators, who indicated on the form they did not know whether they were a translation or a post-editing.

Figure 8. Identification of translations

As for post-edited texts, they were more difficult to identify: while only 22 out of 72 post-edited texts (30.55%) were correctly identified as such, 27 documents were misidentified as translations (37.5%), and in 23 cases (31.94%), evaluators could not be sure of the text-types.

Summing up, although most translations are correctly identified, it seems that post-edited texts are difficult to identify as such, as the great majority of them are either misidentified or not recognized. These results may imply that the quality of post-edited texts can be considered comparable to the quality of translations. However, it remains to be seen to what extent the lack of experience of the evaluators with post-editing may have influenced the results and whether an explicitly mentioned revision task made by the same translators after the translation task would increase the quality of the translations. It should be noticed that translators were instructed to provide a translation that would be fit for recording; hence an implicit revision task was included but not verbalized in the instructions.

5. Discussion of results: correction based on the evaluation matrix

This section analyses in detail the assessments made by evaluators, focusing on the number and type of mistakes found in the translations and post-edited texts, both globally and separately for each excerpt.

Results show that translations needed in general a lower number of corrections than post-editings. The mean difference is 5 errors (see Figure 9), which indicates that the quality of translations and post-edited texts of wildlife documentary scripts could be considered similar. However, when the corrections are analysed separately for each excerpt, it can be observed that there is a considerable difference between the translation quality and the post-editing quality of the first excerpt (see Figure 7), as post-editing almost doubled the mean number of errors of the translation. As far as the second excerpt is concerned, the mean number of corrections in translations and post-edited texts is narrowed (see Figure 9).

Figure 9. Number of corrections

According to the global analysis, the mean number of errors within every issue type is similar (see Figure 10), with 1.195 corrections (2.2. style) the widest difference.

Figure 10. Global evaluation: error typology (both excerpts)

Figure 10 indicates that no corrections were performed on issue type 4.2. (synchrony). Furthermore, it shows that post-edited texts contain more errors in all issue types except in 1.2. (omission), 1.3. (addition), and 2.4. (spelling). Nevertheless, as has been observed in previous analyses, there is a change of tendency in the second excerpt. In this case, translations contain as many or more corrections in 8 issue types: 1.1. (wrong translation), 1.2. (omission), 1.3. (addition), 2.3. (inconsistencies), 2.4. (spelling), 2.6. (grammar), 5 (design/layout) and 6 (others).

Figure 11. Specific evaluation: error typology (excerpt 2)

The greatest difference between the number of corrections in translations and post- editings is found in issue type 2.5. (typography), being 1.5277 points. Compared to the global analysis, excerpt two has two issue types that contain as many corrections (3.583 and 0.389) for translations as for post-edited texts: 1.1. (wrong translation) and 1.2 (omission). In the case of the first excerpt, results are presented in Figure 12.

Figure 12. Specific evaluation: error typology (excerpt 1)

Although results are similar to those of the global analysis, the difference between the corrections performed in issue types 1.1. (wrong translation), 2.2. (style), 2.5 (typography) and 2.6. (grammar) is much wider (2.383, 2.233, 2.133 and 1.250 points respectively) in favor of translation. Moreover, five issue types contain more corrections for translation than post-editing: 1.2 (omission), 1.3 (addition), 2.1 (register), 2.4 (spelling) and 6 (others). No issue type has as many corrections for translating as for post-editing.

On the one hand, results indicate that the errors contained in the translations are more varied, as the subcategories with more errors are from both the categories of accuracy and fluency, as well as domain specific issues: they contain more incorrectly translated words (issue type 1.1) than post-edited texts and more problems regarding register (type 2.1), typography (type 2.5), spotting (type 4.1) and phonetic transcriptions (type 4.3). On the other, results show that post-edited texts usually present more errors in style (type 2.2), grammar (type 2.6) and typography (type 2.5). Moreover, post-edited texts present fewer domain-specific errors in the domain of wildlife documentary films. Thus, it can be observed that, as in other studies that assess post-edited texts versus translations (see Section 2), the quality of post-edited texts is lower only with regards to fluency, which indicates that MT might help with accuracy issues and might allow translators to focus on domain specific issues. Furthermore, it also leads us to believe that better results for the post-edited texts are likely to be obtained through using an MT engine built with in-domain data instead of a free online MT engine like Google Translate, as other errors might be avoided and more terminology could be included.

6. Summary and conclusions

This study is built upon the hypothesis that the quality of post-edited and translated wildlife documentary films is significantly similar and proves its null-hypothesis. The results are presented both globally and separately for each excerpt according to the various types of evaluation data obtained: three scoring rounds, questionnaire-based evaluation, post-editing/translation task identification, and evaluation matrix-based assessment.

The results of all evaluation systems, both globally and separately for each excerpt, correlate and prove that there are no significant differences between post-editing and translation in terms of quality, hence validating the null-hypothesis of the study. Although non-significant, it must be stressed that the differences between translations and post-edited texts vary depending on the excerpt: while translation achieves better results in the first fragment, post-editing has higher marks in the second. Such differences between excerpts might be due to slight differences in their complexity.

When analysing the results of the questionnaire responses, it can be observed that post-edited texts are generally assessed more positively for terminology coherence and domain specific issues, whilst translations are graded better for fluency and general vocabulary. This might indicate that, as post-editors have accuracy issue types solved, they can focus on other issue types, such as domain specific problems and terminological coherence of the text.

As for the correction of the documents based on the evaluation matrix, the results show that the most common errors in texts translated by humans differ from those in post-edited texts: while post- edited texts have many errors regarding style and grammar, translations have more errors regarding mistranslated words. Thus, correction results correlate with the results of the questionnaire, as the most common errors produced in post-edited texts fall into the fluency category, and the most common errors found in translations fall into the accuracy and domain specific categories. As mentioned before, such results might be due to the fact that MT helps with accuracy issues and might allow translators to focus on domain specific issues.

Finally, the results of the translation/post-editing identification task show that evaluators are only able to identify a third of the post-edited texts. If post-edited texts were expected to be significantly different in terms of quality, one could expect a higher number of post-edited texts to be identified as such, but this is not the case. This compels us to claim once again that the quality of translations and post-edited texts in our experiment can be considered similar. However, it is not clear whether the experience of evaluators in terms of post-editing might have influenced such results.

All in all, the results seem to indicate that there are no significant differences between the quality of post-edited texts and translations of the wildlife documentaries used in our experiment. Results also show that the quality of both the translations and the post-edited texts was considered to be low, as the highest mean grade is just a few points above 5, the minimum pass grade in the Spanish system. It remains to be seen whether greater differences would be found in higher quality outputs.

Further research with other language pairs and a higher number of judges should be carried out to confirm the results, because the study is limited to one language pair (English into Spanish) and six human judges. Furthermore, it would be interesting to do similar testing with translations and post-editing produced by experienced translators and post-editors, since in our experiment translators and post-editors were volunteer MA students who had almost no previous professional experience and, consequently, the overall quality was affected by this. Another step would be to research audience reception; in other words, to test how TV audiences would receive a translated wildlife documentary versus a post-edited wildlife documentary. Many possibilities emerge, but this article has hopefully been another step towards future studies on the implementation of translation technologies in the field of audiovisual translation and media accessibility, an area that is still under-researched especially where oral modalities such as voice-over or dubbing are concerned.

References

Armstrong, S., Caffrey, C., Flanagan, M., Kenny, D., & Way, A. (2006). Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation. In *MuTra-Multidimensional Translation Conference Proceedings. Audiovisual Translation Scenarios* (no page numbers). Copenhagen: MuTra. Retrieved from http://www.mt-archive.info/Aslib-2006-Armstrong.pdf [Consulted: 29 June 2016].

Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S., & Sadler, L. (1994). *Machine Translation: An Introductory Guide.* (p. 240). Oxford: NCC Blackwell.

Avramidis, E., Burchardt, A., Federmann, C., Popovic, M., Tscherwinka, C., & Torres, D. V. (2012). Involving Language Professionals in the Evaluation of Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 1127-1130). Istanbul: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/294_Paper.pdf [Consulted: 13 July 2016]

Aziz, W., de Sousa, S., & Specia, L. (2012). PET: a Tool for Post-Editing and Assessing Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (p. 3982-3987). Istanbul: ELRA.

Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf [Consulted: 29th June 2016]

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. (2011). The process of postediting: a pilot study. In *Proceedings of the 8th international NLPSC workshop* (p. 131-142). Frederiksberg: Samfundslitteratur. Copenhagen Studies in Language.

Carroll, J.B. (1966). An Experiment in Evaluating the Quality of Translation. *Mechanical Translation*, *9*(3-4), 55-66.

Dabbadie, M., Hartley, A., King, M., Miller, K.J., El Hadi, W.M., Popescu-Belis, A., Reeder, F., & Vanni, M. (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. In *Proceedings of the Workshop at the LREC 2002 Conference: Machine Translation Evaluation: Human Evaluators Meet Automated Metrics* (8, p. 8-16). Las Palmas: ELRA. Retrieved from https://www.semanticscholar.org/paper/Workshop-at-the-Lrec-2002-Conference-Machine-Palmas/7a4d645f83068c16c919289791c7fe7759677765/pdf [Consulted: 13 July 2016]

Etchegoyhen, T., Bywood, L., Georgakopoulou, P., Fishel, M., Jiang, J., Loenhout, G., Pozo, A., Turner, A., Volk, M., & Maucec, M. (2014). Machine Translation for Subtitling: A Large-scale Evaluation. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)* (p. 46-53). Reykjavik: ELRA. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/463_Paper.pdf [Consulted: 13 July 2016]

Federmann, C., Melero, M., Pecina, P., & van Genabith, J. (2012). Towards Optimal Choice Selection for Improved Hybrid Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, *97*(1), 5–22. Retrieved from http://www.degruyter.com/view/j/pralin.2012.97.issue--1/v10108-012-0001-1/v10108-012-0001-1.xml [Consulted: 13 July 2016]

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

Franco, E., Matamala, A., & Orero, P. (2010). *Voice-over Translation: an Overview* (p. 249). Bern: Peter Lang.

Gambier, Y. (2008). Recent developments and challenges. D. Chiaro, C. Heiss, C. Bucaria (eds). *Between text and image: Updating research in screen translation* (*78*, p. 11-34). Amsterdam/Philadelphia: John Benjamins.

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

Giménez Linares, J.Á. (2008). *Empirical machine translation and its evaluation*. (PhD Dissertation). Universitat Politècnica de Catalunya.

Guerberof Arenas, A. (2009). Productivity and quality in MT post-editing. In *Proceedings of the MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT* (no pages). Ottawa: AMTA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.575.5398&rep=rep1&type=pdf [Consulted: 13 July 2016]

—. (2012). *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. (PhD Dissertation). Universitat Rovira i Virgili.

House, J. (2006). Text and context in translation. *Journal of Pragmatics*, *38*(3), 338-358.

Hutchins, J., & Somers, H. (1992). *An Introduction to Machine Translation* (p. 362). London: Academic Press.

King, M., Popescu-Belis, A., & Hovy, E. (2003). FEMTI: Creating and Using a Framework for MT Evaluation. In *Proceedings of Machine Translation Summit IX* (p. 224-231). New Orleans: AMTA.

Koller, W. (1987). Zum Gegenstand der Übersetzungstheorien. In R. Arntz, & G. Thome (eds). *Übersetzungswissenschaft. Ergebnisse und perspektiven* (p. 19-30). Tübingen: Narr.

Krings, H.P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*, 5. Kent: Kent State University Press.

Kupsch-Losereit, S. (1988). Die Übersetzung als soziale Praxis. Ihre Abhängigkeit vom Sinn- und Bedeutungshorizont des Rezipienten. *Fremdsprachen Lehren und Lernen*, *17*, 203-216.

Lavie, A., & Denkowski, M.J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105-115.

Lommel, A. (ed). (2014) *Multidimensional Quality Metrics (MQM) Specifications*. Retriebed from <http://www.qt21.eu/mqm-definition/mqm-spec-2014-02-14.html> [Consulted: 29 June 2016]

Matamala, A., Fernández-Torné, A., & Ortiz-Boix, C. (2012). Technology and AD: The TECNACC Project. In *Languages and the Media 2012 Conference*. Berlin: ICWE. Retrieved from http://ddd.uab.cat/pub/presentacions/2012/117159/fernandez_matamala_ortiz_berlin2012.pdf [Consulted: 16 May 2016]

Nida, E.A. (1964). *Toward a Science of Translation: with special reference to principles and procedures involved in Bible translating* (p. 331). Leiden: Brill.

Nida, E.A. & Taber, C.R. (1969). *The Theory and practice of Translation* (p. 229). Leiden: Brill.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, *19*(1), 37-58.

Ortiz-Boix, C., & Matamala, A. (forthcoming). Post-Editing Wildlife Documentary Films: a new possible scenario? *Journal of Specialized Translation (JosTrans), 26*, 187-210.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (p. 311-318). Stroudsburg: ACL. Retreived from http://delivery.acm.org/10.1145/1080000/1073135/p311-papineni.pdf?ip=176.85.32.252&id=1073135&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437

&CFID=643378208&CFTOKEN=84140615&__acm__=1468451020_756ba83ab9f75e6b1d06a593338750ed [Consulted: 13 July 2016]

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

Popović, M., & Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, *37*(4), 657-688.

Reiss, K. (1978). Übersetzungstheorie und Praxis der Übersetzungskritik. In F. Königs (ed). *Übersetzungswissenschaft und Fremdsprachenunterricht* (p. 71-93). München: Goethe-Institut.

Reiss, K., and Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie* (p. 254). Tübingen: Niewmeyer.

Roturier, J. (2006). *An Investigation into the impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical.* (PhD Disseratation). Dublin City University.

Savoy, T. (1968). *The Art of Translation* (p. 322). Boston: The Writer.

Snover, M., Dorr, B.J., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas* (p. 223-231). Cambridge: AMTA.

Snover, M., Madnani, N., Dorr, B.J., & Schwartz, R. (2009). Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (p. 259-268). Athens: ACL.

Sousa, S., Aziz, W., & Specia, L. (2011). Assessing the Post-editing Effort for Automatic and Semi-automatic Translations of DVD subtitles. In *Proceedings of Recent Advances in Natural Language Processing Conference* (p. 97-103). Hissar: ACL. Retrieved from http://aclweb.org/anthology/R11-1014 [Consulted: 29 June 2016].

Sutter, N., & Depraetere, I. (2012). Post-edited translation quality, edit distance and fluency scores: report on a case study. Paper presented at the *Journée d'études.*

Toury, G. (1985). *In Search of a Theory of Translation*. Tel Aviv: Tel Aviv University.

Uszkoreit, H., & Lommel, A. (2013). *Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment*. Retrieved from http://www.qt21.eu/launchpad /sites/default/files/MQM.pdf [Consulted: 29 June 2016].

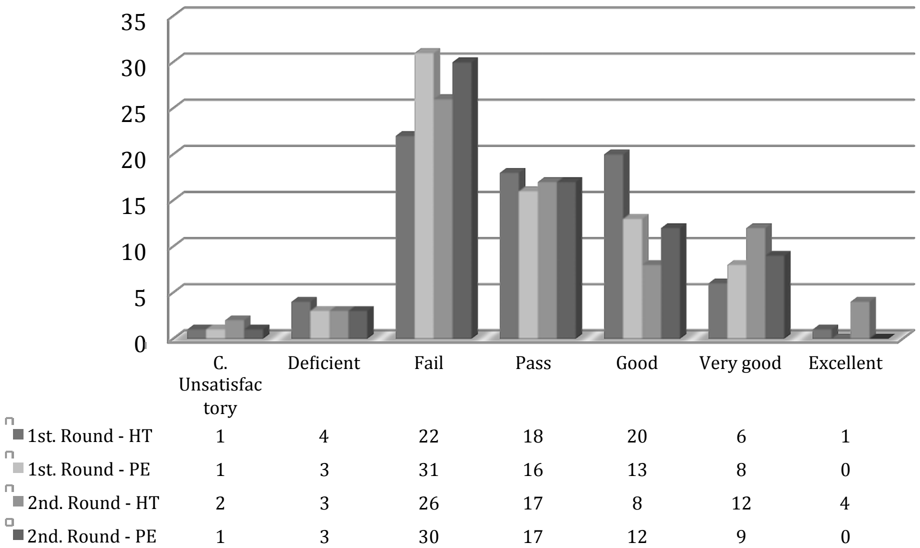
Table 1. Comparison of excerpts

	1st Excerpt	2nd Excerpt
DURATION	101 seconds (1:41 minutes)	112 seconds (1:52 minutes)
WORDS	283	287
ENTRIES/LINES OF SPEECH	8	9
BLEU	44.97	33.75
h-BLEU	51.18	39.5
TER	69.17	59.68
h-TER	74.46	65.34

Table 2. Evaluation matrix: error typology

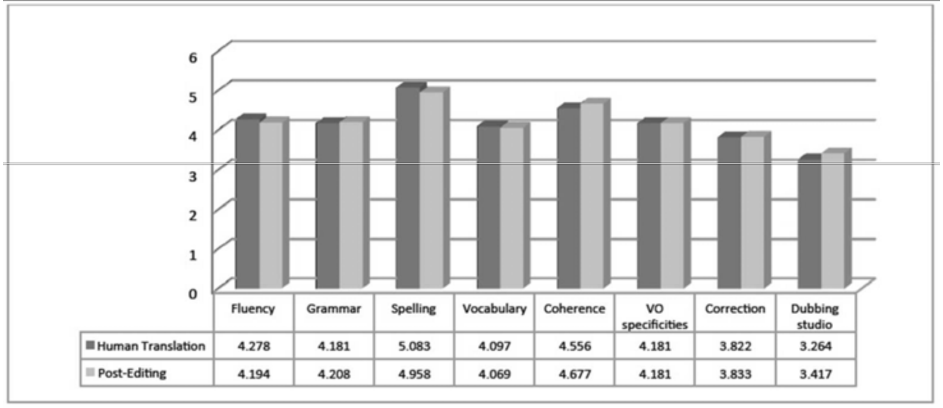
Issue Types Categories	Issue Types Subcategories	ID. Numbers
Adequacy	Wrong translation	1.1.
	Omission	1.2.
	Addition	1.3.
	Non-translated words	1.4.
Fluency	Register	2.1.
	Style	2.2.
	Inconsistencies	2.3.
	Spelling	2.4.
	Typography	2.5.
	Grammar	2.6.
	Others	2.7.
Variety		3
Voice-over specificities	Spotting	4.1.
	Action and kinetic synchronies	4.2.
	Phonetic transcriptions	4.3.
	Isochrony	4.4.
Design / Layout		5
Others		6

Figure 1. Global evaluation results: scoring rounds



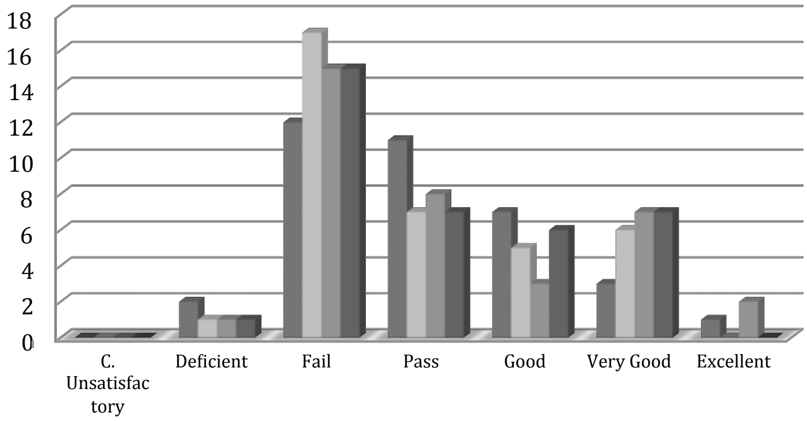
	C. Unsatisfactory	Deficient	Fail	Pass	Good	Very good	Excellent
1st. Round - HT	1	4	22	18	20	6	1
1st. Round - PE	1	3	31	16	13	8	0
2nd. Round - HT	2	3	26	17	8	12	4
2nd. Round - PE	1	3	30	17	12	9	0

Figure 2. Global evaluation results: questionnaire-based assessment (mean values)



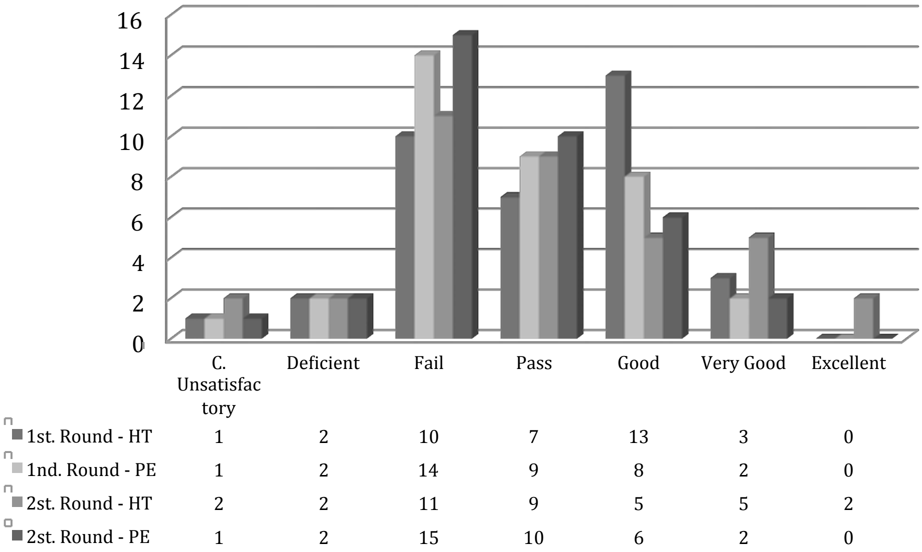
	Fluency	Grammar	Spelling	Vocabulary	Coherence	VO specificities	Correction	Dubbing studio
■ Human Translation	4.278	4.181	5.083	4.097	4.556	4.181	3.822	3.264
□ Post-Editing	4.194	4.208	4.958	4.069	4.677	4.181	3.833	3.417

Figure 3. Specific evaluation results: scoring rounds (excerpt 2)



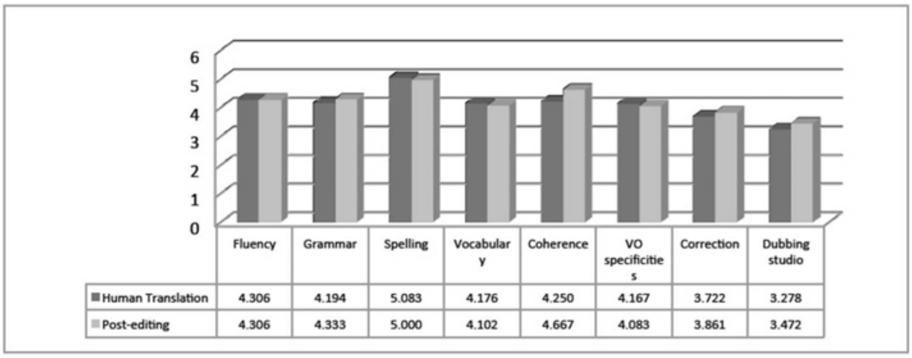
	C. Unsatisfactory	Deficient	Fail	Pass	Good	Very Good	Excellent
■ 1st. Round - HT	0	2	12	11	7	3	1
■ 1nd. Round - PE	0	1	17	7	5	6	0
■ 2st. Round - HT	0	1	15	8	3	7	2
■ 2st. Round - PE	0	1	15	7	6	7	0

Figure 4. Specific evaluation results: scoring rounds (excerpt 1)



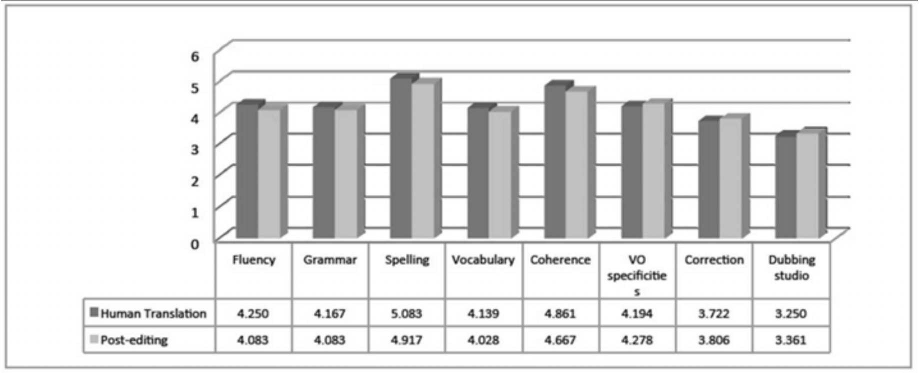
	C. Unsatisfactory	Deficient	Fail	Pass	Good	Very Good	Excellent
■ 1st. Round - HT	1	2	10	7	13	3	0
■ 1nd. Round - PE	1	2	14	9	8	2	0
■ 2st. Round - HT	2	2	11	9	5	5	2
■ 2st. Round - PE	1	2	15	10	6	2	0

Figure 5. Specific evaluation results: questionnaire-based assessment (excerpt 2)



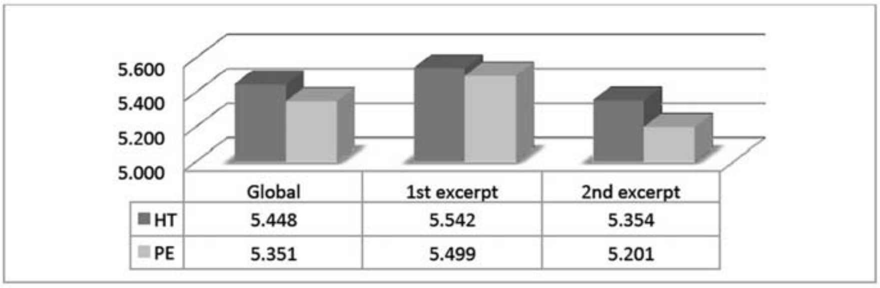
	Fluency	Grammar	Spelling	Vocabular y	Coherence	VO specificitie s	Correction	Dubbing studio
▇ Human Translation	4.306	4.194	5.083	4.176	4.250	4.167	3.722	3.278
▒ Post-editing	4.306	4.333	5.000	4.102	4.667	4.083	3.861	3.472

Figure 6. Specific evaluation results: questionnaire-based assessment (excerpt 1)



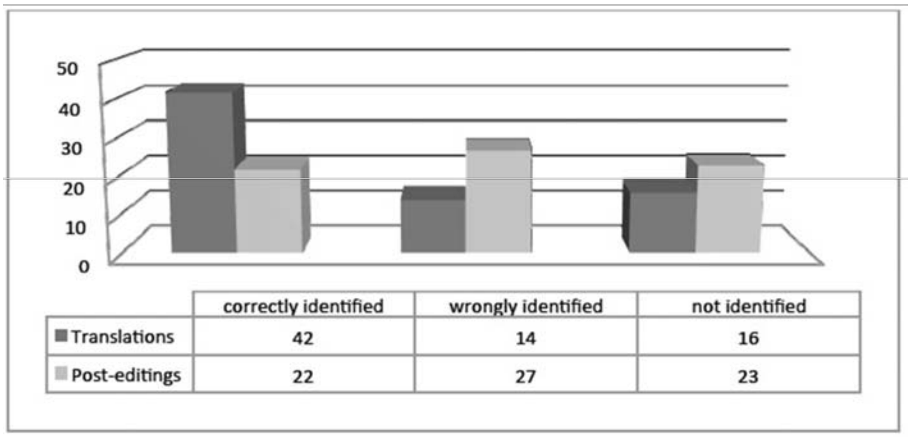
	Fluency	Grammar	Spelling	Vocabulary	Coherence	VO specificitie s	Correction	Dubbing studio
▨ Human Translation	4.250	4.167	5.083	4.139	4.861	4.194	3.722	3.250
□ Post-editing	4.083	4.083	4.917	4.028	4.667	4.278	3.806	3.361

Figure 7. Evaluation round 3



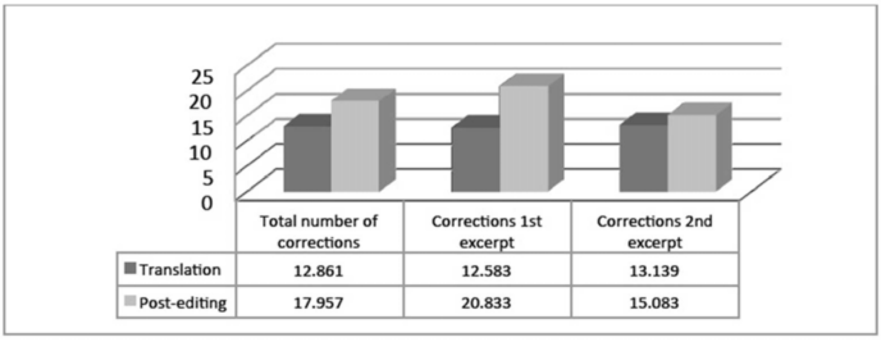
	Global	1st excerpt	2nd excerpt
■ HT	5.448	5.542	5.354
□ PE	5.351	5.499	5.201

Figure 8. Identification of translations



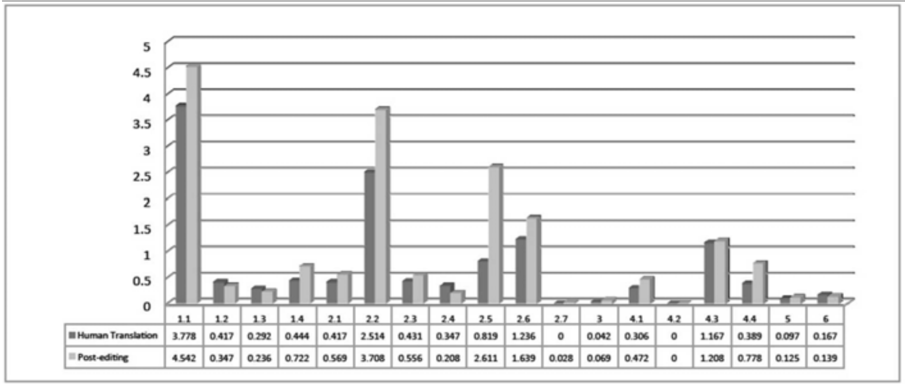
	correctly identified	wrongly identified	not identified
■ Translations	42	14	16
■ Post-editings	22	27	23

Figure 9. Number of corrections



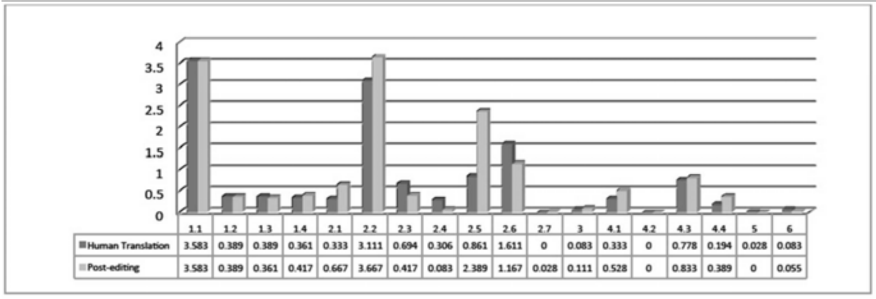
	Total number of corrections	Corrections 1st excerpt	Corrections 2nd excerpt
▆ Translation	12.861	12.583	13.139
▆ Post-editing	17.957	20.833	15.083

Figure 10. Global evaluation: error typology (both excerpts)



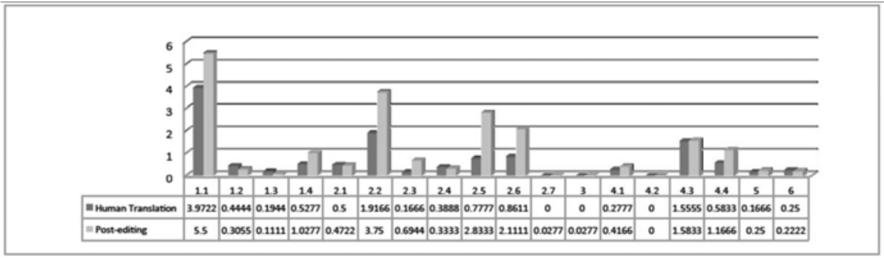
	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3	4.1	4.2	4.3	4.4	5	6
■ Human Translation	3.778	0.417	0.292	0.444	0.417	2.514	0.431	0.347	0.819	1.236	0	0.042	0.306	0	1.167	0.389	0.097	0.167
■ Post-editing	4.542	0.347	0.236	0.722	0.569	3.708	0.556	0.208	2.611	1.639	0.028	0.069	0.472	0	1.208	0.778	0.125	0.139

Figure 11. Specific evaluation: error typology (excerpt 2)



	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3	4.1	4.2	4.3	4.4	5	6
▨ Human Translation	3.583	0.389	0.389	0.361	0.333	3.111	0.694	0.306	0.861	1.611	0	0.083	0.333	0	0.778	0.194	0.028	0.083
▨ Post-editing	3.583	0.389	0.361	0.417	0.667	3.667	0.417	0.083	2.389	1.167	0.028	0.111	0.528	0	0.833	0.389	0	0.055

Figure 12. Specific evaluation: error typology (excerpt 1)



	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6	2.7	3	4.1	4.2	4.3	4.4	5	6
▇ Human Translation	3.9722	0.4444	0.1944	0.5277	0.5	1.9166	0.1666	0.3888	0.7777	0.8611	0	0	0.2777	0	1.5555	0.5833	0.1666	0.25
▇ Post-editing	5.5	0.3055	0.1111	1.0277	0.4722	3.75	0.6944	0.3333	2.8333	2.1111	0.0277	0.0277	0.4166	0	1.5833	1.1666	0.25	0.2222

A4. 4th Article

Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: a Three-Level Approach

Carla Ortiz-Boix carla.ortiz@uab.cat
Anna Matamala anna.matamala@uab.cat
Department of Translation and Interpretation & East Asian Studies, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain

Abstract

This article presents the results of a study designed to evaluate the quality of post-edited wildlife documentary films (in comparison to translated) which are delivered using voice-over and off-screen dubbing. The study proposes a quality assessment at three levels: experts' assessment, dubbing studio's assessment and end-users' assessment. The main contribution of this quality assessment proposal is the inclusion of end-users in the process of assessing the quality of post-edited and translated audiovisual texts. Results show that there is no meaningful difference between the quality of post-edited and translated wildlife documentary films, although translations perform better in certain aspects.

1. Acknowledgements

This article is part of the ALST project (reference FFI-2012-31024) led by Anna Matamala and funded by the Spanish "Ministerio de Economía y Competitividad" and it is also partially funded by research group TransMedia Catalonia (2014SGR27) and the grant FI-DGR2013 awarded by AGAUR. This article is part of the research carried out by Carla Ortiz-Boix under the supervision of Dr. Anna Matamala within the PhD in Translation and Intercultural Studies from the Department of Translation, Interpretation & East Asian Studies at Universitat Autònoma de Barcelona.

2. Introduction

Quality and quality assessment (QA) have been a central issue in Translation Studies since the beginning of the discipline. Many studies have been carried out in that regard (e.g. Nida, 1964; Reiss et al, 1984; Gambier, 1998; Hansen, 2008; Melby et al, 2014), approaching both quality and QA differently depending on the translation theory (House, 2006). Studies on machine translation (MT) and post-editing (PE) have also addressed quality and QA by developing models and measures to evaluate the quality of the text types (technical and general) in which MT and PE is most frequently applied. Although recent studies (Melero et al, 2006; Bywood et al, 2012; Etchegoyhen et al, 2014; Fernández et al, 2013; Ortiz-Boix and Matamala, forthcoming) have proved that including MT and MT plus PE into the workflow of some audiovisual translation (AVT) modalities, mostly subtitling, would positively impact productivity, research into quality and QA of both MT and PE in AVT is still much needed.

This article presents an experiment in which the quality of post-edited wildlife documentary excerpts delivered through voice-over (VO) and off-screen dubbing (OD) has been assessed in comparison to the quality of translations of the same wildlife documentary excerpts. This experiment has been carried out because, after research by Ortiz-Boix and Matamala (forthcoming) demonstrated that applying post-editing instead of translation in these transfer modes could be feasible in terms of effort involved, it is yet to be known how this would impact on quality. Our QA proposal takes into account the specificities of the two audiovisual transfer modes involved (VO and OD) and includes a new aspect that has been usually left aside: the involvement of end-users. It also includes a brief quality assessment by the dubbing professionals that recorded the translated and post-edited versions that were used afterwards in the user reception test.

In order to contextualize our experiment, Section 3 briefly describes the two audiovisual transfer modes under analysis, and summarizes how post-editing QA, and QA in AVT have been approached so far. Section 4 describes the methodological aspects of our QA test. In Section 5, results are presented, and conclusions and further research are discussed in Section 6.

3. Previous Work

This section defines VO and OD, highlighting the specificities of these AVT modalities (3.1). It then summarizes previous work on post-editing QA, with an emphasis on audiovisual translation that has inspired the study (3.2).

3.1. Voice-Over and Off-Screen Dubbing

VO is the AVT transfer mode that revoices an audiovisual text in another target language on top of the source language voice, so that both voices are heard simultaneously (Franco et al, 2010). In countries such as Spain, VO is the transfer mode frequently used in factual programs, e.g. documentary films, as it is said to help reproduce the feeling of reality, truth and authenticity given by the original audiovisual product (Franco et al, 2010). In Eastern Europe, however, VO can also be found in fictional TV programs.

OD is the transfer mode that revoices off-screen narrations substituting the original voice with a version in the target language (Franco et al, 2010). In other words, when OD is applied, only the target language version is heard, not the original one. OD is used in factual programs and usually combined with VO (OD for off-screen narrators, VO for on-screen interviews).

Some of the main features of these transfer modes are the following:

1) Both VO and OD present synchronization constraints. In VO three types of synchrony are observed: kinetic synchrony – the translated text matches the body movements seen on screen–, action synchrony – the translated text matches the actions seen on screen–, and voice-over isochrony – the translated message fits between the beginning and the end of the original speech, leaving some time after the original voice starts and before it ends where only the original can be heard. OD is only endowed with kinetic and action synchronies, as the original voices are not heard in this transfer mode (Orero, 2006; Franco et al, 2010).

2) Different language registers can coexist in audiovisual productions where VO and OD are used: whilst VO is generally used for semi-spontaneous or spontaneous interviews, OD is usually applied to narrators with a planned discourse (Matamala, 2009; Franco et al., 2010). If the original product contains oral features such as fluffs, hesitations and grammatical mistakes, the target language version does not generally reproduce them (Matamala, 2009). In other words, the translation is generally an edited version of the original.

VO and OD are often used to revoice wildlife documentary films from English into Spanish, the object of our research. This type of non-fictional genre usually includes many terms that might pose additional challenges to the translators (Matamala, 2009). It is also often the case that the source text contains linguistic errors and inconsistencies (Franco et al, 2010), and that a quality written script is not available (Ortiz-Boix, forthcoming). However, translators are expected to deliver a quality written script in the target language so that the recording by voice talents in a dubbing studio can begin.

3.2. Post-Editing Quality Assessment

Although research on QA of post-edited text has increased, it is still rather limited. Fiederer and O'Brien (2009), Plitt and Masselot (2010), Carl et al (2011), García (2011), Guerberof (2009, 2012), Melby et al (2014) and Mariana (2014) have dealt with quality in post-editing, to a greater or lesser extent. Up until now, QA has been based mostly on what has been has termed in the QTLauchPad project (Lommel et al, 2014) as either holistic approaches –which assess the quality of the text as a whole – or analytic approaches –which assess the quality by analysing the text in detail according to different sets of specifications. A combination of both can also be found.

Holistic approaches: Plitt and Masselot (2010) used the Autodesk translation QA team to assess randomly selected samples of translated and post-edited text using two labels ("average" or "good"), depending on whether they considered the text was fit for publishing. In Carl et al (2011), raters ranked the quality of a list of sentences, either translated or post-edited. Fiederer and O'Brien (2009) also assessed the quality of sentences – three translated and three post-edited versions of 30 sentences – according to clarity, accuracy and style on a 4-point scale. Raters were also asked to indicate their favorite option out of the six proposals for each source sentence.

Analytic approaches: In García (2011), a rater assessed the quality of a 500-word text by using the Australian National Accreditation Authority for Translators and Interpreter's (NAATI) guidelines. In Guerberof (2009, 2012), three raters blindly assessed translated segments, post-edited segments and segments previously extracted from a translation memory by using the LISA QA model.

Mixed approaches: Melby et al (2014), Mariana (2014) and Lommel et al (2014) develop and implement the Multidimensional Quality Metrics (MQM) in their analysis. The model provides a framework for defining metrics and scores that can be used to assess the quality of human translated, post-edited or machine translated texts. It sets error categories, otherwise called issue types, which assess different aspects of quality and problems. MQM is partly based on the translation specifications (Melby, 2014) that define expectations for a particular type of translation; MQM is organized in a hierarchic tree that can include all the necessary issue types for a given text type and a given set of specifications.

In the specific field of audiovisual translation, post-editing quality assessment research is still more limited: EU-financed project SUMAT (Etchegoyhen et al, 2014) evaluated the quality of the machine translation output via professional subtitlers who assigned a score to each subtitle. They were asked for general feedback on their experience while post-editing as well as on their perceived quality of the output. Aziz et al (2012) assessed the quality of the machine translated subtitles by post-editing them using the PET tool. The post-edited subtitles were afterwards assessed against translated subtitles using BLEU and TER automatic measures, suggesting there is no meaningful difference in terms of quality between them.

4. Methodology

Our experiment involved one language pair (English into Spanish), and aimed to assess the quality of post-edited wildlife documentaries compared to the quality of human translations. It is built upon the hypothesis that there is no meaningful difference between the quality of post-editing and the quality of translation of wildlife documentaries in English delivered through VO and OD in Spanish.

The experiment included a three-level quality assessment: (1) quality assessment by experts, with a mixed approach (holistic and analytic); (2) quality assessment by the dubbing studio where the translations and post-editings were recorded, and (3) quality assessment by end-users, who watched both post-edited and translated audiovisual excerpts. The inclusion of end-users in the assessment has been inspired by functionalist approaches to translation and by recent user reception studies in AVT. In the case of wildlife documentaries, we wanted to assess whether both post-edited and translated documentaries fulfilled their function to the same extent, that of informing and entertaining the audience.

4.1. Participants

Participants taking part on the first level assessment were six lecturers of MAs on audiovisual translation in universities in Spain who are experts on VO and currently work or have recently worked as professional voice-over translators. The experts' profiles are comparable: all of them have a BA in Translation Studies except for one, who has a BA in German Studies. Furthermore, five of them have either a PhD in Translation or have attended PhD courses on the same field. Previous experience varies among participants: when the experiment was carried out experts 1, 3, and 5 had worked as audiovisual translators between 10 and 16 years and taught for 11, 8, and 5 years respectively, while participants 2, 4, and 6 had between 5 and 8 years of experience as audiovisual translators and taught for the last 4 or 5 years. The number of experts used to rate the documents is higher than in previous studies on QA and post-editing (Guerberof, 2009; García, 2011; or De Sutter et al, 2012)

For the second level, only one dubbing studio was used, as only one study was needed to record the materials. Two voice talents, a dubbing director and a sound technician were present during the recording session.

In the third level, 56 users with different educational backgrounds took part in the experiment (28 male, 28 female, 23-65 years old, mean age: 39.15). All participants were native speakers of Spanish and 46.43% of the participants were highly proficient in English. Watching habits related to wildlife documentaries do not vary much among participants (96.43% watch a maximum of 3 documentaries on TV every month), but preferences in terms of the audiovisual transfer mode to be used in wildlife documentaries differ: 30.46% prefer subtitling, 44.64% prefer dubbing, and 25% prefer VO. These preferences are correlated with age: participants under 40 prefer subtitled documentaries (50%), whilst participants over 40 prefer voiced-over documentaries (46.3%).

4.2. Materials

The materials used for the first level were 6 translations and post-editings of two self-contained excerpts of a 7-minute wildlife documentary film titled *Must Watch: a Lioness Adopts a Baby Antelope* that is currently available on Youtube as an independent video (http://www.youtube.com/watch?v=mZw-1BfHFKM). It is part of the episode *Odd Couples* from the series *Unlikely Animal Friends* by National Geographic broadcast in 2009. Short excerpts were chosen for practical reasons, despite being aware that this could impact on evaluative measures of enjoyment and interest. Additionally, excerpts of a wildlife documentary were chosen since documentaries follow structured conventions and have specific features in terms of terminology (Matamala, 2009). The translations and post-editings (24 in total) were produced by 12 students of an MA on AVT that had had a specific course on VO but no, or almost none, previous experience on post-editing. Hence, they were instructed to correct all the errors and adjust, only if necessary, the text according to the specific constrains of documentary translation. Participants worked in a laboratory environment that recreated current working conditions: they used a .doc document and they were allowed to use any available resources (internet, dictionaries, etc.) To perform both tasks, students were given a maximum of 4 hours, although almost none of them used the entirety of the given time. The audiovisual excerpts were similar in terms of length (first excerpt: 101 seconds, 283 words; second excerpt: 112 seconds, 287 words) and content, and the translations and post-editings contained between 218 and 295 words. They were machine translated through *Google Translate*, the best free online MT engine to be used to machine translate wildlife documentary scripts according to Ortiz-Boix (forthcoming).

For the second level, the best post-editing and the best translation of each excerpt was selected, according to the results of the first-level quality assessment. The recordings of these excerpts were used for the third-level assessment.

4.3. Test Development

Level 1: Experts' Assessment. Participants carried out the experiment from their usual place of work. They were given detailed instructions on how to assess the 24 documents without knowing which of them were translated or post-edited. They were given 20 days to perform the whole assessment. The experiment was divided into three evaluation rounds:

a) In round 1, raters were instructed to read each document and grade it according to their first impression on a 7-point scale (completely unsatisfactory-deficient-fail-pass-good-very good-excellent). They were just given one day for this task, and the order of the documents was randomized across participants.

b) In round 2, raters were asked to correct the documents following a specific evaluation matrix (see section 4.4.), and grade them after the correction on a 7-point scale. Afterwards, they had to answer an online questionnaire (see section 4.5.).

c) In round 3, a final mark between 0 and 10, following Spain's traditional marking system, was requested.

There was also a final task in which raters had to guess whether the assessed document was translated or post-edited (post-editing/translation identification task).

Level 2: Dubbing Studio Assessment. The scripts and videos were sent to the dubbing studio and a professional recording was requested from them. They were instructed to follow standard procedures. A researcher took observational notes and gathered quantitative and qualitative data on the changes made during the recording session by the dubbing director.

Level 3: End-Users' Assessment. Quality was understood to be based on end-user reception and, following Gambier's proposal (2009), three aspects were assessed: understanding, enjoyment, and preferences (or response, reaction and repercussion in Gambier's terms). Participants were invited to a lab environment that recreated the conditions in which documentaries can be watched: they sat in an armchair and watched the documentary excerpts in a 32' flat screen. Taking into account ethical procedures approved by Universitat Autònoma de Barcelona's ethical committee, participants were administered a pre-task questionnaire (see section 4.6.). They were then shown two of the excerpts without knowing whether they were watching a translated or post-edited excerpt. After each viewing, a questionnaire was administered to them to test their comprehension and enjoyment, as well as their preferences (see section 4.6).

4.4. Evaluation Matrix (Level 1)

The evaluation matrix applied in the first level is based on MQM because it can be used for both translations and post-editings, and it also allows to select and add only the relevant categories for our text type. Although MQM offers the possibility to include over one hundred issue types, only five categories and eleven subcategories of issue types were selected, as shown on Table 1.

Issue types categories	Issue types subcategories
Adequacy	Wrong Translation
	Omission
	Addition
	Non-translated words
Fluency	Register
	Style
	Inconsistencies
	Spelling
	Typography
	Grammar
	Others
Variety	
Voice-over/off-screen dubbing specificities	Spotting
	Action and kinetic synchronies
	Phonetic transcriptions
	VO Isochrony
Design/Layout	
Others	

Table 1. Evaluation matrix: error typology

The selection was based on previous research on errors produced by MT engines in general texts (Avramidis et al, 2012) and wildlife documentary films (Ortiz-Boix, forthcoming), as well as in post-editings (Guerberof, 2009). As MQM does not contain a domain specific issue type for audiovisual translated texts, a new category was added: VO/DO specificities. It includes the issue types subcategories spotting, action and kinetic synchrony, voice-over isochrony, and incorporation of phonetic transcriptions. Raters were trained on how to apply the evaluation matrix.

4.5. Questionnaire design (Level 1)

The questionnaire in level 1 aimed to gather the agreement of the raters with eight statements assessing fluency, grammar, spelling, vocabulary, terminological coherence, voice-over specifications, satisfaction, and success in terms of purpose, using a 7-point Likert scale:

- In general, the text was fluent.

- In general, the translation was grammatically correct.
- In general, there were no spelling issues.
- In general, the vocabulary was appropriate.
- In general, the terminology was coherent throughout the text.
- In general, the translation met the VO and DO specificities.
- In general, the final result was satisfactory; aka the translation met its purpose.
- In general, the translation could be sent to the dubbing studio to be recorded.

4.6. Questionnaire design (Level 3)

The pre-task questionnaire included five open questions on demographic information (sex, age, highest level of studies achieved, mother tongue, and other spoken languages) as well as seven questions on audiovisual habits.

The post-task questionnaire included seven questions on enjoyment. Participants had to report their level of agreement on a 7-point Likert scale on the following statements:

- I have followed the excerpt actively.
- I have paid more attention to the excerpt than to my own thoughts.
- Hearing the Spanish voice on top of the original English version bothered me.
- I have enjoyed watching the excerpt.

They also had to answer the following questions on a 7-point Likert scale:

- Was the excerpt interesting?
- Will you look for more information regarding the couple presented on the documentary?
- Would you like to watch the whole documentary film?

They were also asked 3 questions on perceived quality and comprehension, again on a 7-point Likert scale:

- The Spanish narration was completely understandable.
- There were expressive problems in the Spanish narration.
- There were mistakes in the Spanish narration.

Five additional open questions per excerpt were used to test comprehension. Finally, participants were asked which excerpt they preferred. A pilot test was run to validate the questionnaire, which was inspired by Gambier (2009).

4.7. Data and Methods

The following data were obtained:

Level 1 (experts):

1) 144 documents with corrections (6x24) according to the MQM-based evaluation matrix.

2) The grades for each document in the three scoring rounds.

3) 144 completed questionnaires (6x24 documents) reporting on the participants' views after correcting each document.

4) The results of the post-editing/translation identification task.

Level 2 (dubbing studio):

5) 4 documents with corrections (1x4) made by the dubbing director and their corresponding recordings.

6) Observational data gathered during the recording session.

Level 3 (end-users):

7) 56 completed questionnaires on demographic aspects and audiovisual habits.

8) 112 completed questionnaire responses (14x4) on user enjoyment, comprehension and preferences. In order to analyse the comprehension questionnaire, wrong answers were given 0 points, partially correct answers were assigned 0.5 points and correct answers, 1 point.

All data were analysed using the statistical system R-3.1.2, developed by John Chambers and colleagues at Bell Laboratories. In this study, data was analysed according to descriptive statistics.

5. Discussion of Results

Results are presented according to the three levels of assessment. More attention is devoted to levels 2 and 3, as a more detailed analysis of the first level is already presented in Ortiz-Boix and Matamala (forthcoming).

5.1. Quality Assessment by experts[1]

The quality of both translations and post-editings was rather low and no meaningful differences between post-editings and translations in terms of quality were found, as the difference between the scores for each of the tasks were low. Results are discussed in two different sub-sections: in the holistic approach, the scores given in the evaluation rounds, the questionnaire replies and the identification task results are analysed. The analytic approach discusses the results of the corrections performed by the raters.

5.1.1 Holistic Approach

Results of round 1 indicate that experts evaluate better translations than post-editings after reading the documents for the first time: while 45 out of 72 (62.5%) translations were evaluated from "pass" to "excellent", only 37 out of 72 post-editings (51.39%) were evaluated within this range. However, when documents are rated again after a thorough correction (round 2), the difference between post-editings and

	Passes for Round 1	**Passes for Round 2**
Translations	**45**	**41**
Post Editings	**37**	**38**
Total Possible	**72**	**72**

Table 2. Pass marks for round and task

[1] See Ortiz-Boix and Matamala (forthcoming) for further information on the results of this level.

translations diminishes. In this case, 41 out of 72 translation (56.94%) and 38 out of 72 post-editings (52.78%) are given between a "pass" and an "excellent". Despite these slight differences, the median grade in both rounds is a “pass” for both translations and post-editings.

Results for round 3, in which the Spanish traditional marking system was used (from 0 to 10, 5 being a “pass”), show again a very small difference: the mean grade for translations is 5.44 versus 5.35 for post-editings. This mark correlates perfectly with grades obtained in rounds 1 and 2.

As for the questionnaire replies, results indicate that post-editings are given higher grades in four of the issue types – grammar, terminological coherence, satisfaction, and success in terms of purpose– and the exact same grade in the case of VO specificities. Translations are considered better in fluency, vocabulary appropriateness, and spelling. However, no relevant differences are found in any case.

Concerning the final identification task, experts correctly categorized 42 translations out of 72 (58.33%) and only 22 post-editings (30.56%). They categorized wrongly 14 translations (19.44%) and 27 post-editings (37.5%), and could not decide whether the document was a translation or a post-editing in the case of 16 translations (22.22%) and 23 post-editings (31.94%). Results indicate that post-editings are more difficult to identify than translations, as the great majority of them are either misidentified or not recognized as such. If the quality of post-editings were generally worse, a better identification would be expected, which leads us to suggest that the quality of both translations and post-editings is comparable.

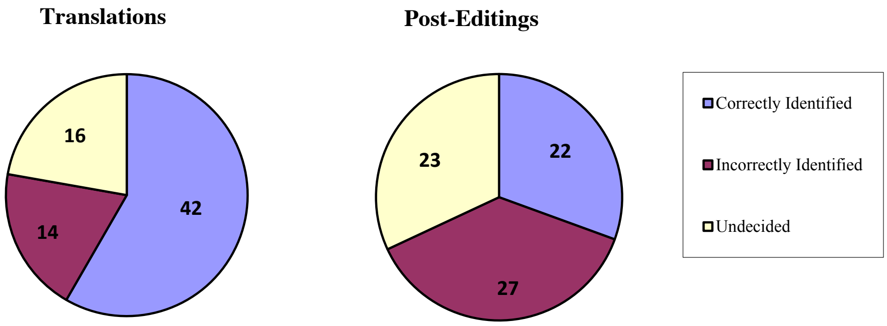


Figure 1. Task identification

5.1.2. Analytic Approach: Correction

Translations present a lower number of corrections (mean: 12.861 per document) than post-editings (17.957), although the mean difference in a text is five corrections and it is not meaningful. It is interesting to highlight that experts did not correct any errors regarding synchrony and did a higher number of corrections for post-editings in all issue types but three: omission, addition, and spelling (see Ortiz-Boix and Matamala forthcoming for further details). The issue types with more errors, both in post-editing and translation, were wrong translation, style, typography, and grammar. Given the small differences, results seem to prove that the quality of both translations and post-editings in our experiment was similar, although the type of errors found in either post-editings or translations differ.

5.2. Quality Assessment by the Dubbing Studio

During the recording session it was observed that changes made in the translation and post-editing scripts were only related to aspects directly linked to the voicing of the documentaries.

In the first excerpt, a similar number of changes were made: six in the post-editing excerpt, five in the translated excerpt. Changes referred to synchronization aspects (3 in the translated version and 2 in the post-edited one), phonetics (2 and 1 respectively), and stylistic repetitions (0 and 3 respectively); the experts in level 1 had surprisingly not corrected issues related to synchronization. In the second excerpt, 4 changes were made in the translated version (1 on phonetics and 3 on synchronization). As for the post-editing, the dubbing director pointed out that the synchronization was not good and that a re-translation was needed. However, it was decided to record it as it was and test whether audiences would react negatively.

Although no quantitative differences were observed, data show that the translation, at least in the second excerpt, was qualitatively better than the post-edited script.

5.3. Quality Assessment by Users

Data were analysed taking into account all participants but a more specific analysis divided participants into two age groups (group A: <40, group B: >40) as differences in terms of preferences for subtitling or VO were observed in the demographic questionnaire. Results are presented in terms of enjoyment and preferences (see section 5.3.1.) and understanding (see section 5.3.2.).

5.3.1. End-Users Enjoyment and Preferences

Results indicate that, regardless of the excerpt, version, and age group, users mostly agree with the fact that they followed the excerpt actively (median for all conditions/groups/excerpts = “strongly agree”) and focused on what they were watching on screen (all medians are “strongly agree”, except for post-editing of excerpt 1= “moderately agree”). Hearing the Spanish voice on top of the original English version did not bother any of the participants in any of the conditions or excerpts (median = “strongly disagree” with the statement “Hearing the Spanish voice on top of the original English version bothered me”), although percentages show a difference between age groups: older viewers (96.43%) are not bothered at all by the Spanish voice on top of the original English voice (“strongly disagree” with the statement), whilst the percentage in younger viewers drops (57.14%). This percentage, though, is distributed evenly across both versions, showing that it is the transfer mode (VO) and not the translation system (translation/post-editing) that impacts on them. This also correlates with the preferences stated by younger audiences in the pre-task questionnaire.

	Excerpt 1		**Excerpt 2**	
	Translation	**Post-editing**	**Translation**	**Post-editing**
I have followed the excerpt actively	Strongly agree	Strongly agree	Strongly agree	Strongly agree
I have paid more attention to the excerpt than to my own thoughts	Strongly agree	Moderately agree	Strongly agree	Strongly agree
Hearing the Spanish voice on top of the original English version bothered me	Strongly disagree	Strongly disagree	Strongly disagree	Strongly disagree
I have enjoyed watching the excerpt	Strongly agree	Moderately agree	Moderately agree	Strongly agree

Table 3. Agreement level on enjoyment (medians)

When asked to express their level of agreement or disagreement with the statement “I have enjoyed watching the excerpt”, users grade the translated version higher than the post-editing one (translation median= “strongly agree”, post-editing median: “moderately agree”). Although there are slight differences depending on the excerpt: in excerpt 1, 57.14% of the participants strongly agree with the statement whilst the percentage drops to 32.14% in the post-editing, being the median “strongly agree” for the translation and “moderately agree” for the post-editing. In excerpt 2, differences in enjoyment are higher: 85.71% of the users who watched the post-edited version strongly or moderately agree with the statement, in contrast with 57.14% of the users of the translated version. The median for the post-editing is “strongly agree” and for the translation it is “moderately agree”. Slight differences are observed between age groups, since overall the younger group “moderately agrees” with the statement and the older group “strongly agrees”, but no differences are found between translations and post-editings within each group.

Apart from enjoyment, one direct question (“Was the excerpt interesting?”) with seven different options (from “very interesting” to “very boring”) aimed to assess their interest in the film. Overall results show that the translation was better evaluated than the post-editing (“translation median = “very interesting”, post-editing median= “pretty interesting”), although differences are found in the two excerpts under analysis: in excerpt 1 the translation is considered by all participants as either “very” or “pretty interesting”, whilst the post-editing is only considered as “very” or “pretty interesting” by 67.87% of participants. It is even qualified as “boring” by 10.71% of the participants. The difference is minimal though, as the median in both cases is "pretty interesting" for excerpt one. In the second excerpt, the trend changes: 82.14% consider the translation “very” or “pretty interesting”, whilst 100% qualify the post-editing as such. The difference in this case is higher, as the median is “very interesting” for post-editings and “pretty interesting” for translations. These are unexpected results since the dubbing studio

		Was the excerpt interesting?
Excerpt 1	**Translation**	Pretty interesting
	Post-editing	Pretty interesting
Excerpt 2	**Translation**	Pretty interesting
	Post-editing	Very interesting

Table 4. Agreement level on interest (medians)

considered the second excerpt post-editing to be of low quality. When analysing the data according to the age groups, it can be observed that the 40 and over group prefer the translation (85.71% rated it as "pretty interesting" and 14.29% as "very interesting" while the younger group like the post-editing better (100% rated it as "very interesting"). To gather more information on interest, participants were also asked whether they would be willing to look for more information on the documentary, and the median reply in all conditions, regardless of excerpt, age and condition, was "maybe" (the middle option on a 7-point Likert scale). Similarly, to the question "Would you like to watch the whole documentary film?", a positive reply was obtained in all conditions (median= "yes"), regardless of age. The only difference is found in the second excerpt, where those who watched the translated version react more positively (median= "yes") than those that watch the post-edited (median = "maybe").

Finally, when asked which of the two versions was their preferred one, without knowing which one was a post-editing or a translation, results show almost no difference between both versions: while 44.64% of the participants prefer a translated version, 42.86% prefer a post-edited one. However, when excerpts are analysed separately, it can be seen that participants prefer the translated version (50%) to the post-edited (35.71%) for the first excerpt, and the post-edited (50%) to the translated (39.26%) for the second. Differentiating between age groups, older viewers prefer the translated versions of both excerpts to the same extent (85.71%), whereas younger viewers prefer the post-edited version of the second excerpt (85.71%) and the translated version of the first (78.57%).

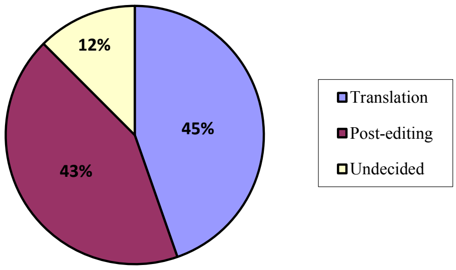


Figure 2. Preferred versions

Overall results show slightly better results in some aspects for the translation (enjoyment, interest, and preferences), although different trends are observed when analysing the data independently for excerpts and age groups.

5.3.2. End-Users Comprehension

All participants considered the narration to be completely understandable and did not perceive any mistakes. However, results show slight differences in comprehension in some instances. Taking into account both excerpts and all participants, translated versions are better understood (mean score: 0.71) than post-edited ones (mean score: 0.66). When analysing each excerpt separately, opposite trends are observed: the translation is better understood in the first excerpt (translation= 0.79, post-editing= 0.63), whilst the post-editing is slightly better understood than translation in the second one (translation= 0.63, post-editing= 0.69). Considering both age groups, the younger group seems to understand better translated versions (translation= 0.72, post-editing= 0.61), whilst the older group obtains almost identical results (translation= 0.70, post-editing= 0.71).

In conclusion, results show slightly higher comprehension levels for the translation when considering all the data. Translation is also slightly higher in comprehension for the first excerpt and the younger group. Almost identical results are found for the older group, and slightly higher results in favor of post-editing are encountered for the second excerpt.

6. Conclusion and Further Research

This article presents an experiment in which the quality of post-edited wildlife documentary films is compared to the quality of translated documentaries in order to determine whether there is a meaningful difference between the qualities of each. Compared to other QA performed in the field of Translation Studies and PE, the QA used in this experiment was carried out in three levels: it takes into account not only experts but also end-users and the dubbing studio where the written script is converted into an oral recording.

The results of the study indicate that, according to experts, translations seem to perform better in the three evaluation rounds when global percentages are considered, but median results show no differences. A lower number of corrections is also performed on translations, although the differences are low. On the contrary, post-editings are better graded in more aspects than translations in the questionnaire after round 1, although differences are again minimal. And, finally, post-editings are more difficult to identify as such, which may be considered an indicator that no meaningful quality differences are observed.

When observational data from the dubbing session are analysed, translation also seems to perform better, although the differences in the first excerpt are minimal and higher in the second one.

Finally, when taking into account end-users, better median results are obtained for the translation in terms of enjoyment, interest, and user preferences, although a meticulous analysis of each excerpt and group yields diverging trends. It must be stressed, though, that the differences are low, and the same results are obtained for both conditions in the other items under analysis. In terms of comprehension, translation is better understood than post-editing when taking into account all the data, but also in the first excerpt and in the younger group. However, results are non-meaningful.

All in all, translation seems to receive better marks, although the difference is not high, and hence, not meaningful, proving our initial hypothesis.

When comparing the evaluation at the three stages, it can be inferred that expectations of end-users are not high, as their ratings were high compared to the rather low evaluations of both experts and the dubbing studio professionals. The low quality of both translations and post-editings might be due to the lack of experience of the MA students and the test conditions (volunteer work rather than professionally paid commission), which is a limitation of our research. It remains to be seen whether professional translators, with or without post-editing experience, would yield different results.

This study is limited in scope but it hopefully will open the door to future research in the field of audiovisual translation evaluation and post-editing. Future studies could take into account other language pairs, work with longer excerpts, and involve professional translators as well as experts in post-editing. Another stakeholder could be included in the evaluation, namely the broadcaster commissioning the VO of non-fictional genres. It may well be that quality expectations, and consequently evaluations of lecturers, professionals, broadcasters, dubbing directors, and end-users differ in many aspects, and analysing these different expectations is an interesting research topic. Additionally, a modified version of our experiment could include methodological improvements such as developing identical questions at different levels in order to obtain comparable data. We are fully aware that our research can be improved and expanded in many ways, but it has hopefully contributed to shed some light on an under-researched topic.

References

Avramidis, E., Burchardt, A., Federmann, C., Popovic, M., Tscherwinka, C., and Vilar, D. (2012). Involving Language Professionals in the Evaluation of Machine Translation. *LREC*, 1127-1130.

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. *LREC*, 3982-3987.

Bywood, L., Volk, M., Fishel, M., and Georgakopoulou, P. (2013). Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives*, *21*(4), 595-610.

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. 2011. The process of postediting: a pilot study. *Proceedings of the 8th international NLPSC workshop*. Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds). (Copenhagen Studies in Language 41), Frederiksberg: Samfundslitteratur: 131-142.

De Sutter, N., Depraetere, I. (2012) Post-edited translation quality, edit distance and fluency scores: report on a case study. Proceedings, *Journée d'études Traduction et qualité Méthodologies en matière d'assurance qualité*, Université Lille 3, Sciences humaines et sociales, Lille.

Etchegoyhen, T., Fishel, M., Jiang, J., and Maucec, M. S. (2013). SMT Approaches for Commercial Translation of Subtitles. *Machine Translation Summit XIV, Main Conference Proceedings*, 369-370.

Fernández, A., Matamala, A., and Ortiz-Boix, C. (2013). Enhancing sensorial and linguistic accessibility with technology: further developments in the TECNACC and ASLT projects. *5th International Media For All Conference*. *Audiovisual Translation: Expanding Borders*. Dubrovnik, 25-27 September 2013.

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

Franco, E., Matamala, A., and Orero, P. (2010). *Voice-over translation: An overview*. Peter Lang.

Gambier, Y. (1998). *Translating for the Media*. University of Turku.

Gambier, Y. (2008). Recent developments and challenges. *Between text and image: Updating research in screen translation* 78: 11.

García, I. 2011. Translating by post-editing: Is it the way forward? *MachineTranslation*, Vol. 25(3). Netherlands: Springer. 217-237

Guerberof, A. (2009). Productivity and quality in MT post-editing. *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*.

Guerberof, A. (2012). *Productivity and quality in the post-editing of utputs from translation memories and machine translation*. . Universitat Rovira i Virgili.

Hansen, G. (2009). The speck in your brother's eye–the beam in your own. *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile* 80 (2008): 255.

House, J. (2006). Text and context in translation. *Journal of Pragmatics*, *38*(3), 338-358.

Lommel, A., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, *12*:455-463.

Mariana, V. R. (2014). The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment, Brigham Young University MA Thesis.

Matamala, A. (2009) Main Challenges in the Translation of Documentaries. *New Trends in Audiovisual Translation*. Ed. Díaz Cintas, Jorge. Bristol: Multilingual Matters. Chapter 8.: 109-120

Melby, A., Fields, P., Koby, G. S., Lommel, A., and Hague, D. R. (2014). Defining the Landscape of Translation. In *Tradumàtica* (pp. 0392-403).

Melero, M., Oliver, A., and Badia, T. (2006). Automatic Multilingual Subtitling in the eTITLE project. *Proceedings of ASLIB Translating and the Computer 28*. London.

Nida, E. A. (1964). *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Orero, P. (2006) Voice-over: A case of hyper-reality. *EU High Level Scientific Conference Series. MUTRA Proceedings*.

Ortiz-Boix, C. (Forthcoming). Post-Editing Wildlife Documentaries: Challenges and Possible Solutions.

Ortiz-Boix, C. and Matamala, A. (forthcoming). Post-Editing Wildlife Documentary Films: a new possible scenario?

Ortiz-Boix, C. and Matamala, A. (forthcoming). Assessing the Quality of Wildlife Documentaries.

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

Reiß, K., & Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie* (Vol. 147). Walter de Gruyter.