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.

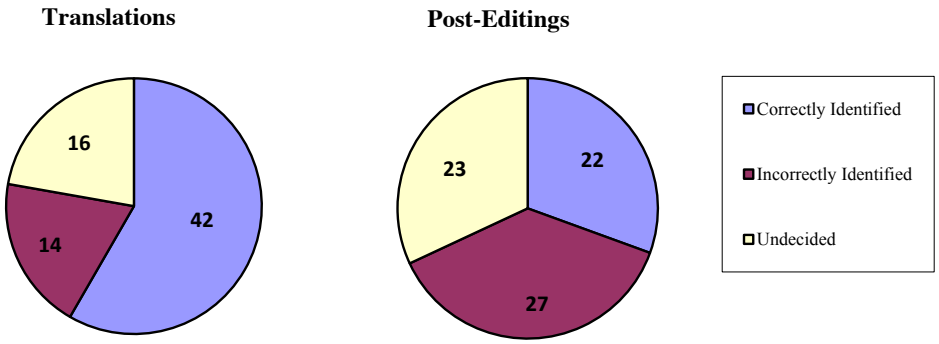

Figure 1. Task identification

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%).

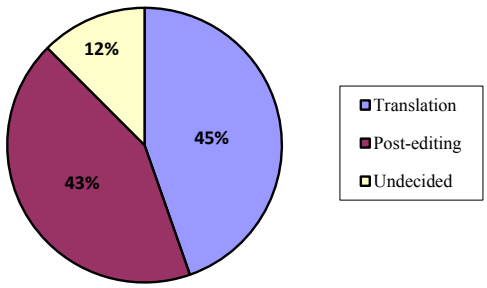

Figure 2. Preferred versions

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.