Machine translation in audio description? Comparing creation, translation and post-editing efforts

Anna Fernández-Torné

Anna Matamala

Abstract: Machine translation has been proved worthwhile, in terms of time saving and productivity gains, in technical and administrative translation domains. In order to examine whether this also applies to audio description, an experiment comparing the efforts of creating an audio description from scratch, of translating it manually from English into Catalan and of post-editing its machine translated version has shown that the objective post-editing effort is lower than creating it ex novo. *However, the subjective effort is perceived to be higher.*

Keywords: accessibility; audio description; audiovisual translation; machine translation; Catalan language; post-editing effort

Introduction

The presence of audiovisual content in our society is increasing at a dramatic pace. New ways of making this growing volume of audiovisual content accessible to all audiences faster – and at lower costs, if possible – need to be researched and developed, and the implementation of technologies in audiovisual translation (AVT) seems to be the way forward, as it has already been proved efficient in other translation domains.

Machine translation (MT) is one of the technologies that is becoming common practice in the professional translation arena (Koponen 2015, Daems et al. 2015), and translators' productivity gains using MT have been broadly demonstrated (Guerberof 2009; Plitt and Masselot 2010). MT with post-editing (PE) – that is, with a revision by a professional – is already part of the workflow of many translation service providers dealing with technical texts and also of public administrations aiming "to quickly check the general meaning of incoming information" (European Commission n.d.). However, "[t]he adoption rate of MT and PE processes naturally varies in different countries and language pairs" (Koponen 2015: 3), and in translation domains, too. This is where audiovisual translation in general, and audio description in particular, lags behind. Audio descriptions, the translation of images into words addressed to an audience who cannot access the visual content (Maszerowska et al. 2014), are nowadays generally created independently in each language and are only seldom translated, being the application of MT being non-existent to the best of our knowledge.

This article presents the results of an experiment in which MT was implemented in audio description (AD) for the English-Catalan language pair. The experiment compared the effort, both objective and subjective, in three different scenarios: when creating an audio description in Catalan (AD creation), when translating an English audio description into Catalan (AD translation), and when post-editing a machine-translated audio description from English into Catalan (AD PE). Our ultimate aim is to explore whether MT could be satisfactorily deployed in audio description, hence the focus of the analysis is the comparison of AD PE in relation to AD creation, which is currently the standard process. However, another possibility has also been taken into account, i.e. human translation, a process already discussed in the literature in relation to audio description (Matamala 2006, Jankowska 2013). Results in this regard are also provided, although they are discussed to a lesser extent.

The article begins with an overview of related work. Next, the experimental set-up is presented, with a thorough description of the participants, test data, effort assessment methods, test development, and statistical methods used. In the following section, a comprehensive exposition of the results is presented and discussed, and finally, conclusions are drawn while proposing directions for further research.

Related work

The application of MT to audiovisual content is still in its early stages. In recent decades the EU has funded several projects dealing with the automatic generation of subtitles and their translation into multiple languages both in media – MUSA (2002-2004), eTITLE (2003-2005), SUMAT (2011-2014) and EU-BRIDGE (2012-2014) –, and educational content – transLectures (2011-2014) and EMMA (2014-2016). Research has also been carried out to assess the quality of machine-translated or post-edited audiovisual translations such as subtitles (Armstrong et al. 2006; Volk 2009; Del Pozo et al. 2014) and, more recently, voice-overs (Ortiz-Boix and Matamala 2015). However, the implementation of MT in audio description has not yet attracted the attention of many researchers, and only the ALST project (Matamala 2015) has ventured into the topic, proving so far the feasibility of machine translating filmic AD in the Catalan-Spanish language pair (Ortiz-Boix 2012). This article is part of that project, and focuses on comparing the effort involved in generating an AD when using different methods. That is why this section will succinctly describe previous research in post-editing effort, placing special emphasis on its measurement.

The general framework used in many studies to assess post-editing effort is Krings' (2001) proposal. Krings differentiates between temporal, technical and cognitive effort. Temporal effort is the total time spent on post-editing a text, technical effort refers to the operations carried out to post-edit the text, and cognitive effort applies to the mental processes involved in identifying errors in raw machine-translated texts and in deciding on the necessary steps to correct them.

Measuring temporal effort is straightforward. Technical effort can also be directly observed by using methods such as key-logging technologies (Guerra 2003, Tatsumi and Roturier 2010). However, cognitive effort is not directly observable. Krings (2001) used think-aloud protocols to determine cognitive effort, but he noticed that this method affected the total process time. Other technologies, such as key-logging (O'Brien 2004) and eye-tracking (O'Brien 2011, Carl et al. 2011), have successfully been used, since they allow subjects' behaviour to be recorded unobtrusively in real time. Pauses have been considered a key indicator of cognitive effort. Indeed, in writing research pauses are "assumed to provide us with a window to the cognitive processes underlying language production" (Wengelin 2006: 108, cited in Chukharev-Hudilainen 2014: 64), and they are usually computed, particularly their frequency, duration and position. Lacruz, Denkowski and Lavie (2014) state that both average pause ratio (APR), i.e. the average time per pause divided into the average time per word, and pause to word ratio (PWR), i.e. the number of pauses divided into the number of words, correlate well with cognitive effort: the lower the APR and the higher the PWR, the higher the levels of cognitive effort.

Most of the research carried out so far in this area has focused on technical documents, since this is where machine translation is more extensively used. In the field of audiovisual translation, studies on post-editing effort are more limited: De Sousa, Aziz and Specia (2011) compare the temporal effort involved in translating subtitles from English into Brazilian Portuguese compared to post-editing draft versions produced using translation tools, both MT and TM. Results show that "translating from scratch consistently takes 70% longer than post-editing the same sentence" (*ibid.*: 5). On the other hand, Ortiz-Boix and Matamala (forthcoming) compare the effort involved in translating wildlife documentary excerpts compared to post-editing them. Their results seem to indicate that post-editing may imply less effort than translating, although statistically significant results are not achieved in all parameters under analysis.

Methodology

This section describes methodological aspects such as the selection of participants, the test data, the measurement tools, the test development, and the statistical methods. The whole procedure was approved by the Ethics Committee of the Universitat Autònoma de Barcelona.

Participants

The participants' profile was controlled to avoid high variability which could distort the results of the test. Volunteers were recruited from among native Catalan-speaking students of an MA in AVT.

Fourteen participants took part in the experiment, but for technical reasons only the results of twelve could be used. It should also be noted that one task of one participant was not adequately recorded (translation of clip A), but since the other data were available they were included in the analysis.

Two participants were male (17%), and ten were female (83%), with a mean age of 25.8 years. All but one had a BA in Translation and Interpreting and all of them finished their MA in Audiovisual Translation in June 2014, when the test took place. They had the same experience as far as AVT and AD creation was concerned: only as students had they translated audiovisual products and created ADs.

In relation to their attitude towards translating ADs and of post-editing machine-translated ADs, participants showed a general negative prejudice towards post-editing machine-translated ADs. Prior to the test, when presented with the statement "Machine translating ADs created in other languages and post-editing them conveniently is useful" and asked to express their level of agreement on a 5-point Likert scale (1 being "strongly disagree" and 5, "strongly agree"), two participants (16.6%) chose 1, six participants (50%) chose 2, and four participants (33.3%) chose 3. When the statement presented was "Translating ADs created in other languages is useful", only one participant selected 2 (8.3%), seven selected 3 (58.3%) and four participants (33.3%) selected 4, indicating a more positive attitude towards human translation. When asked to comment on their choices, they argued that MT plus PE would lack naturalness, would convey more calques, and the task itself would often be as time-consuming as creating an AD from scratch.

Test data selection

Three clips from the film *Closer* (2004, directed by Mike Nichols) were chosen as test data. This film was selected for various reasons. First, since this experiment was part of a wider project in which other technologies such as speech recognition (SR) (Delgado, Matamala and Serrano 2015) and text-to-speech (Fernández-Torné and Matamala 2015) were tested in AD, a film both in English and in Catalan (dubbed version), with AD in both languages, was required. Secondly, a film with a non-specific genre addressed to adults was favoured, so that children films were considered out of scope, and a film within a 'miscellaneous' category according to the classification by Salway et al. (2004) was searched for.

The clips' duration was established at approximately three minutes to minimise participants' fatigue, as they would have to create, translate and post-edit three different AD excerpts in just one session. The number of words included in the AD to be translated was also controlled to balance the test duration, so that in no case would the translation and post-editing take longer than one hour. Neutral clips in terms of content were chosen in order to avoid any potential distraction or offense to the participants. Finally, clip excerpts from the development of the plot, rather than the beginnings, were chosen as specific constraints are generally to be found in terms of AD creation at the beginning of a film (Remael and Vercauteren 2007).

For each clip three versions were available: (1) for the AD creation, the audiovisual Catalan dubbed version of the excerpts; (2) for the human translation task, the audiovisual Catalan dubbed version with the audio description in English provided as written text with time codes. This AD corresponds to the one included in the commercial DVD released in 2005; (3) for the AD PE task, the audiovisual Catalan dubbed version with the audio description in English, provided as written text with time codes, plus the machine translation generated by Google Translate of the English AD, also provided as written text with time codes. Google Translate was chosen as the best free online engine available in the chosen language pair and domain in a pre-test (Fernández-Torné forthcoming).

Assessment measures

Following Krings (2001), effort was split up into three categories: temporal effort, technical effort, and cognitive effort. Even though this classification was designed for the assessment of effort in post-editing tasks, it was also deemed adequate for evaluating creation and translation efforts, since they can all be considered comparable indicators of text production, as explained by Dam-Jensen and Heine (2013). According to the authors, there are three types of text production, i.e. writing, translation and adaptation, which relate differently to pre-existing texts. In this sense, adaptation – post-editing in our case – "can be seen as an 'intermediate type' as it depends on a source text (or more than one), as does translation, but involves a shift in text type by means of paraphrasing, revising or summarizing" (Dam-Jensen and Heine 2013: 92).

Although "post-editing time, a simple and objective annotation, can reliably indicate translation post-editing effort in a practical, task-based scenario" (Specia 2011: 73) and can also be seen "as a way to assess some of the cognitive effort involved in post-editing" (Koponen et al. 2012: 1), effort was assessed by measuring several parameters from each category, namely:

- Temporal effort: total process time, time spent in Subtitle Workshop.
- Technical effort: keyboard actions (including total character types and other keystrokes), mouse actions (including clicks, movements and scrolls), switches keyboard to mouse, and total window transitions.

- Cognitive effort: total pause time, mean pause time, number of pauses and PWR.

All these elements were automatically recorded by the key-logging tool InputLog 5.2.01 (Leijten and Van Waes 2013), and are referred to in the analysis as objective effort.

The total process time was measured to determine the temporal effort. An additional indicator was the time spent in the software where the actual creation, translation or post-editing took place: our belief is that the less the time spent in Subtitle Workshop, the more temporal effort involved in searching for information or solving doubts on the Internet.

Regarding the technical effort, both keyboard actions (itemising total characters typed and other keystrokes) and mouse actions (differentiating between clicks, movements and scrolls) were calculated. Although deletion and insertion operations are considered to be direct indicators of technical effort (Krings 2001: 179), they could not be recorded in the selected software. Instead, switches from keyboard to mouse and total number of window transitions were computed, as these are also operations made during the process.

Concerning cognitive effort, PWR was calculated as the main indicator of cognitive effort in post-editing (Lacruz, Denkowski and Lavie 2014). Other aspects related to pauses – total pause time, number of pauses, mean pause time – were also assessed, as pauses have been found to be good indicators of cognitive demand, not only in writing research but also in translation (Lacruz, Shreve and Angelone 2012). It must be highlighted at this point that O'Brien (2006) did not find significant evidence to prove that pauses are actually related to cognitive effort in post-editing, but since they have largely been proved to correlate well with cognitive load in both written and spoken language production and translation research, they were considered in the present study, where they are defined as any scriptural inactivity of more than 300 ms (Lacruz, Denkowski and Lavie 2014).

Apart from these objective measures, it was considered interesting to assess the participants' subjective effort, similar to what De Sousa, Aziz and Specia (2011) did. Data on participants' perceived effort and opinions were gathered via a questionnaire administered after each task, and was compared to the participants' expected effort and opinions, gathered also via a questionnaire.

Questionnaire design

A profile questionnaire (PQ) was designed to gather personal information on participants, such as age, sex, and level of education.

A general questionnaire (GQ) was developed to gather the participants' attitudes to post-editing and translating audio descriptions, and their opinions on various aspects both before performing the test (expectations) and after having performed it (perceptions). The GQ included four statements for each of the tasks under analysis (AD creation, AD translation, AD PE) to which participants had to indicate their level of agreement on a 10-point numerical scale:

- Rate the tasks according to the effort you think they will involve for you
- Rate the tasks according to how much you think they will impair creativity
- Rate the task according to how much you think they will be boring
- Rate the task according to the quality you think they will achieve

A slight variation was included in the GQ to be administered after the experiments: verb tenses were changed from "will involve" to "have involved", and an additional open field to justify their choices was added.

As can be seen from the previous statements, the issues under analysis relate to effort, creativity impairment, boredom, calque conveyance, and output quality, as these are aspects often mentioned in relation to post-editing. Subjective ratings were deemed important not only to complement objective data, but also to check whether their expectations on the tasks were met and to examine whether their attitudes towards any of the tasks changed once they had performed them.

Three post-task questionnaires (PTQ) were also designed to obtain data on the participants' views immediately after performing each of the tasks. A first set of questions asked participants to rate their level of agreement with a series of statements on a 5-point Likert scale, including an open field for comments. The statements read:

a) In the AD creation PTQ:
   - The clip was easy to audio describe.

b) In the AD translation PTQ:
   - The source text was easy to translate.
   - The clip was easy to audio describe departing from the original AD.

c) In the AD PE PTQ:
   - The clip was easy to audio describe departing from the MT AD.
   - The machine-translated text was easy to post-edit.
   - The machine-translated text required no post-editing.
   - The machine-translated text was fluent Catalan.
   - All the information in the source text was present in the machine-translated text.

Additionally, in the AD translation and in the AD PE PTQ, a question specifically asked whether there were any elements participants had had to adapt from the departure text (be it the English AD or the MT output) and, if so, which. Possible answers included "amount of information", "length of descriptions, "frequency of descriptions", "number of incomplete sentences (with no verb)", "register (too formal or too colloquial"), and also an open field.

As can be seen from the previous statements, some of them allowed for an easy comparison between tasks, for instance in terms of ease.

Test development

The experiment was carried out in a controlled environment (laboratory conditions), following a within-subjects design. A pilot test allowed improvements to the experimental design.

The experiment was divided into two parts. In the first part participants were asked to fill in the PQ and the GQ. They were then requested to watch the Catalan dubbed version of the film *Closer* from beginning to end uninterruptedly, so that they all had the same contextual information. Then there was a 30-minute break.

In the second part of the experiment, they were asked to create the Catalan AD, to translate the English AD into Catalan and to fully post-edit the English to Catalan machine-translated AD of the three three-minute-long excerpts.

The instructions for the AD creation stated that they should deliver a Catalan audio description according to the Catalan AD style. As for the AD translation, they were told that an English AD with spotting (time coding of the AD units) would be given to them and their task was to create a Catalan AD, modifying time-codes and AD units if needed. They were told that they should adapt the original AD to the Catalan AD style, which should fit with the Catalan dubbed version provided. The same instructions were used for the AD PE task. Moreover, the following specific guidelines inspired by the works of O'Brien (2010), TAUS and CNGL (2010), Specia (2011), De Sousa, Aziz, and Specia (2011) and Housley (2012) were included for PE:

- Perform the minimum amount of editing necessary to make the AD translation ready for voicing retaining as much raw translation as possible
- Aim for grammatically, syntactically and semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- Ensure that the message transferred is accurate.
- Ensure that key terminology is correctly translated.
- Basic rules regarding spelling, punctuation and hyphenation apply.

The order of the tasks and clips was balanced across participants. Participants were asked to perform all three tasks using Subtitle Workshop 2.51 (http://subworkshop.sourceforge.net/index), a software they were all familiar with. Although it is a subtitling software, Subtitle Workshop was chosen because it includes an integrated video player and allows inserting or editing time codes where appropriate for the synchronisation of the audio description.

After performing each task, a PTQ was administered to all participants. Once all tasks were finished, they were asked to complete the GQ, as described in the previous sub-section.

Statistical methods

Descriptive statistics (mean, median, standard deviation, minimum and maximum) were computed for all quantitative variables. A bivariate analysis was performed to determine the relationship between each variable and the task being performed. For the comparison of the tasks, a repeated measures model was used, taking into account that each participant had performed all three tasks. All results were obtained using SAS, v9.3 (SAS Institute Inc., Cary, NC, USA). For the decisions, significance level was fixed at 0.05.

Results and discussion

This section presents and discusses the results in the three tasks under analysis: AD creation, AD translation, and AD PE. Objective effort results are presented first, followed by the analysis of subjective effort and participants' views. When differences between tasks are statistically significant, it is explicitly mentioned in the discussion. Non-statistically significant data are also provided because they may illustrate relevant differences in the processes.

Objective effort

Temporal effort

Mean total process times for the AD creation and AD PE tasks were quite close to each other: 2,696.880 seconds (44.95 minutes) was the mean total process time for AD creation, whereas 2,666.695 seconds (44.44 minutes) was the total for the AD PE task. Although the figure for AD translation was higher (2,919.641 seconds, i.e. 48.66 minutes), there were no statistically significant differences among the three tasks.

The amount of time spent in Subtitle Workshop, where the actual task was to be performed, was also calculated. AD PE and AD translation presented a closer mean time (2,238.552 seconds, i.e. 37.31 minutes, and 2,245.303 seconds, i.e. 37.42 minutes, respectively) spent on the software, and for AD creation the time spent was only slightly higher (2,415.218 seconds, 40.25). Again, the difference was not statistically significant.

When calculating relative values (see *Figure 1*), it was observed that in AD creation participants spent 90% of the time in Subtitle Workshop, which means it was the task requiring less research on the Internet, whereas AD translation was the task requiring most time outside Subtitle Workshop (33%). Post-editing was somewhere in between, dedicating 84% to the Subtitle Workshop and 16% to searching the Internet. These results can be seen as a logical consequence of the processes associated with each task: while AD can be considered a creative and introspective task, translation is usually associated with dictionary searches and on-line consultations. On the other hand, PE mainly implies rewording, word reordering and error correction, which do not necessarily involve as many Internet searches.

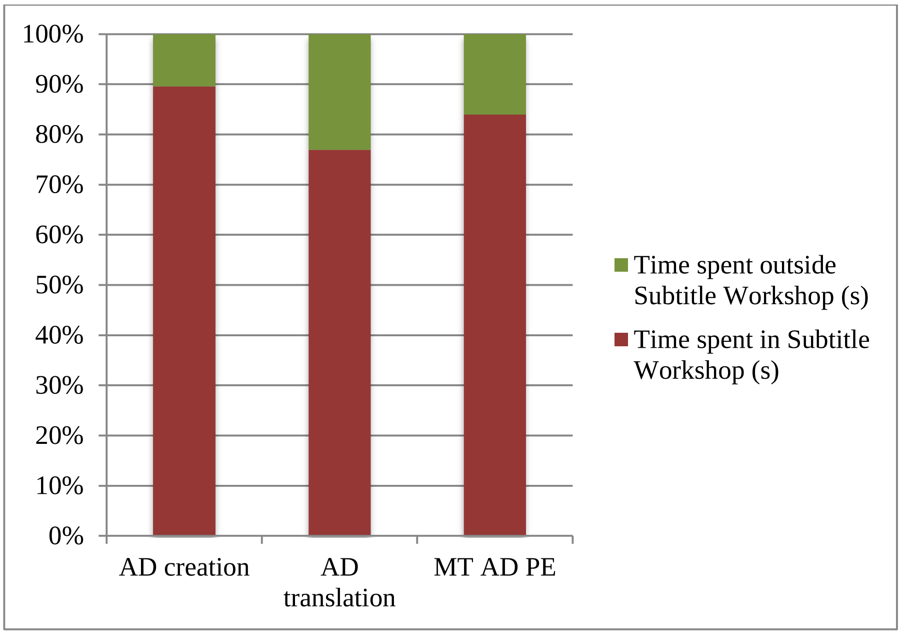

Figure 1 Time spent inside and outside Subtitle Workshop

Globally, although differences were not statistically significant, AD post-editing was the task that presented the lowest total process time and therefore, the least temporal effort.

Technical effort

Taking all keyboard actions as a whole, AD creation rendered the highest number of keyboard actions, with an average of 2,948.417. AD translation had an average of 2,656.545 actions, while AD PE presented only 1,973 actions on average. However, only the difference between AD creation and AD PE was statistically significant.

When only the total number of characters typed (including spaces) was taken into consideration, AD PE showed a significantly lower number of characters than the other two tasks: a mean total number of 885.083 characters typed against 1,520 for AD creation and 1,763.727 for AD translation. As for the rest of keystrokes, AD translation showed the lowest number of keystrokes on average, with only 892.818, followed by AD PE (1,087.917) and AD creation (1,428.417). Even though the results for AD creation were higher, there were no statistically significant differences between any of the tasks.

Mouse actions presented quite similar mean figures: 1,473.667 for AD creation, 1,556.583 for AD PE and 1,666.545 for AD translation. In this respect, clicks and movements did not show significant differences either, but scrolls did (see *Table 1*). AD creation (23.417 scrolls on average) was statistically lower than both AD translation (65 scrolls) and post-editing (58.667 scrolls).

Concerning the number of switches from keyboard to mouse, all means ranged from 209 to 232, showing no statistically significant difference. It was in the total number of window transitions that significant differences were to be found again: AD translation presented a statistically higher number of transitions (209.727) than AD creation (99.167), but not AD PE (141). This result is in line with the distribution of time spent inside and outside Subtitle Workshop: AD creation was the task which spent proportionally more time in Subtitle Workshop and it was also the task showing the lowest amount of transitions, with the post-editing task falling between AD creation and AD translation.

Globally, in relation to technical effort, post-editing was statistically the least keyboard intensive task, with significantly the lowest number of characters typed, in accordance with O'Brien's (2010) findings. It was also the task entailing fewer mouse clicks and fewer switches from keyboard to mouse, while the rest of the values were not the highest for the three tasks in any case. All this seems to indicate that post-editing is the task involving less technical effort.

Cognitive effort

Concerning the mean total pause time, post-editing showed the lowest mean total pause time (1,394.345 seconds, i.e. 23.24 minutes), followed by AD translation (1,504.525 seconds, i.e. 25.08 minutes) and AD creation (1,625.437 seconds, i.e. 27.09 minutes). AD PE also presented the lowest mean number of pauses (961.083), although both AD creation and AD translation were not far away from that figure, presenting a very similar mean number of pauses (1,031.583 and 1,035.091, respectively). The mean time of such pauses did not differ much either: while AD PE presented the lowest mean pause time (1.505 seconds), AD translation had a mean pause time of 1.514 seconds and AD creation, of 1.724 seconds. No statistically significant differences were found in any of these items.

In connection with the pause to word ratio (PWR), AD PE showed a statistically lower mean ratio (4.081) than AD creation (6.009), but not AD translation (4.591).

It was deemed interesting to see whether the distribution between the time spent pausing and the time devoted to active writing diverged from task to task. AD creation seemed to be assigning more time to pauses (60.27%), while AD translation and post-editing devoted just a little more than half of the time to pausing (51.53% and 52.29% respectively) (see *Figure 2*). Even though the difference was not significant, it is important to highlight that the task of creation involves more pausing than writing, which might be an indicator of a higher cognitive effort.

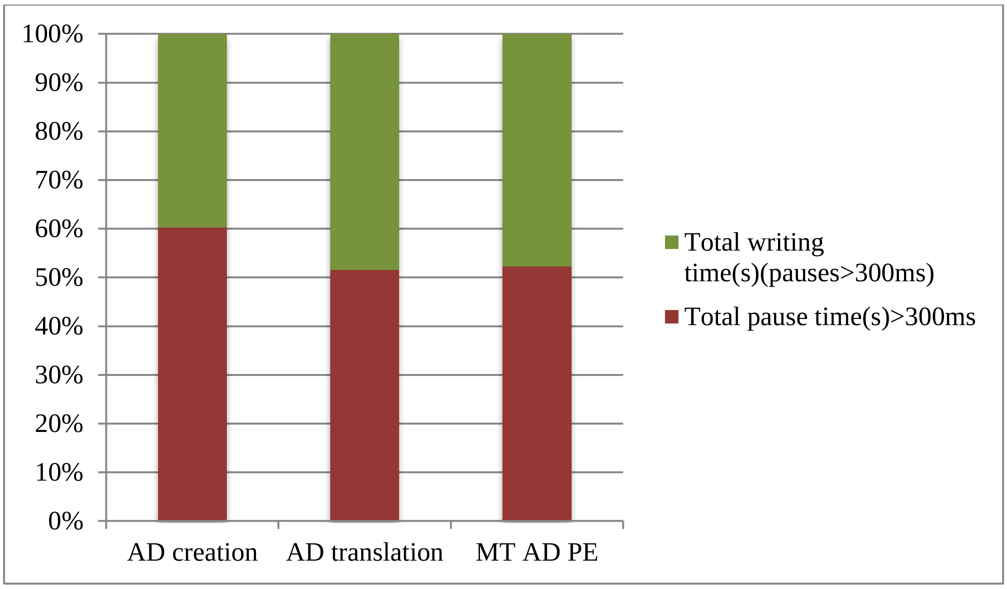

Figure 2 Distribution of pausing and writing during each task

All these data seem to indicate that post-editing was the least effort-involving task, especially if we focus on a key indicator such as PWR: AD PE presented the lowest number of pauses and the highest number of words, resulting in the lowest PWR, which is associated with low levels of cognitive effort. Conversely, AD creation seems to be the most demanding cognitively. *Table 1* presents an overview of objective results.

		AD creation	AD translation	AD post-editing
Temporal effort	total process time (seconds)	2,696.880	2,919.641	2,666.695
		(44.95 minutes)	(48.66 minutes)	(44.44 minutes)
	time spent in Subtitle Workshop (seconds)	2,415.218	2,245.303	2,238.552

		(40.25 minutes)	(37.42 minutes)	(37.31 minutes)
Technical effort	keyboard actions	2,948.417	2,656.545	1,973.000
	total number of characters typed (including spaces)	1,520.000	1,763.727	885.083
	other keystrokes	1,428.417	892.818	1,087.917
	mouse actions	1,473.667	1,666.545	1,556.583
	left clicks	615.333	616.364	567.000
	right and middle clicks	3.833	4.727	2.500
	movements	831.083	980.455	928.417
	scrolls	23.417	65.000	58.667
	switches from keyboard to mouse	231.333	223.182	209.583
	window transitions	99.167	209.727	141.000
Cognitive effort	total pause time (seconds)	1,625.437	1,504.525	1,394.345
		(27.09 minutes)	(25.08 minutes)	(23.24 minutes
	number of pauses	1,031.583	1,035.091	961.083
	mean pause time (seconds)	1.724	1.514	1.505
	pause to word ratio	6.009	4.591	4.081

Table 1 Overview of objective effort assessment results

Subjective effort and participants' opinions

Beyond objective effort, this research aimed to go a step further and gather data on participants' subjective views on effort and other relevant aspects. First of all, a comparison of the replies to the GQ, before and after the experiment, is presented, focusing first on effort and then on other items such as the degree of creativity impairment each task involves, boredom, calque conveyance, and final quality. Secondly, the participants' opinions after each task are analysed, adopting a contrastive approach where possible.

General questionnaire responses

Quantitative data on the participants' expected effort (prior to the tasks) and perceived effort (after the task) was gathered through a questionnaire, which included also an open field to justify their choices in its post-task version. Opinions on other aspects were also gathered. *Table 2* shows the means and medians obtained for each item under analysis before and after performing the tasks, on a 10-point scale where 1 is the lowest value. In the case of final quality, however, it must be clarified that "best quality" was number 1 whilst "worst quality" was number 10.

		AD creation		AD translation		AD post-editing	
		Pre	Post	Pre	Post	Pre	Post
Effort involved	Mean	8.25	7.17	6.17	5.58	6.50	7.50
	Median	8	7	6	6	6	8
Creativity impairment	Mean	3.09	3.82	7.45	7.27	8.45	9.36
	Median	3	4	8	7	9	10
Boredom	Mean	2.09	1.82	4.18	4.18	6.73	7.27
	Median	2	2	4	4	6	8
Calque conveyance	Mean	1.25	2.00	5.25	5.42	6.93	8.33
	Median	1	1.5	5	5	7	9
Final quality	Mean	1.67	2.58	2.75	3.25	4.83	5.08
	Median	2	2	2.5	3	4.5	5

Table 2 Comparison of opinions before and after the experiment

Results indicate that participants expected AD PE to be the task that would impair their creativity the most and would convey more calques. They also expected it to be the most boring task, and the one delivering the worst output quality, and involving more effort. However, in terms of effort, once the experiment was finished, both AD creation and translation were perceived as involving less effort than expected (mean=8.25 and 6.17 prior to the test to 7.17 and 5.58 after the test), while PE AD showed the opposite trend (6.50

changed into 7.50), becoming the task involving more effort according to our sample of participants. Regarding the other indicators, they all showed a clear evolution towards worse PE ratings after performing the task. This was also the case for most indicators in AD creation and AD translation, except for creativity impairment in AD translation (7.45 prior to the test, 7.27 after), and boredom in both AD translation (4.18 both prior and after the test) and AD creation (2.09 into 1.82). One possible explanation for this trend is the lack of experience of our participants.

As regards open questions that provide qualitative data, the fact that time codes were already given to participants both in the translation and post-editing tasks was often stressed as an advantage as far as effort was concerned, but the poor quality of the machine-translated text was seen as a drawback since “while a few sentences were translated correctly, most of them had mistakes or the structure needed changes” (Participant 3). Qualitative answers also reinforced the idea of post-editing being the most creativity-impairing task as it imposes “a constraint to the final text” (Participant 7). However, some participants pointed to the instruction indicating them to keep as much raw MT text as possible as the reason behind this creativity impairment rather than the actual usage of MT, which comes to show the importance and impact of instructions not only in the research arena but also in the professional world.

In connection with the degree of boredom of the tasks, responses reasserted that “[t]he AD creation task is the least boring task” (Participant 5) since “the more creative you can be, the less boring the activity will be” (Participant 7), which seems to indicate they enjoyed it more, although enjoyment was not directly assessed in the questionnaire. They also agreed in terms of conveying calques that “[i]n both the translation and the MT AD post-editing you risk to use [*sic*] calques because you do not create a new text, but depart from a source text in a foreign language” (Participant 9), and that “MT AD lacks quality because the audiodescriber [*sic*] departs from a text which is not perfectly translated” (Participant 9).

On the basis of the above, it seems that post-editing was the task involving the most subjective effort of all and presenting more drawbacks, which contrasts with objective data analysed previously.

Post-task questionnaires analysis

One of the questions included in all three post-task questionnaires assessed how easy participants felt a particular task was, immediately after performing it. Although not explicitly mentioning effort in the statement, this measure can somehow be linked to the effort participants perceived in the task. *Figure 3* shows how many participants selected a specific value on a 5-point Likert scale for each task, 1 being “strongly disagree” and 5 being “strongly agree”.

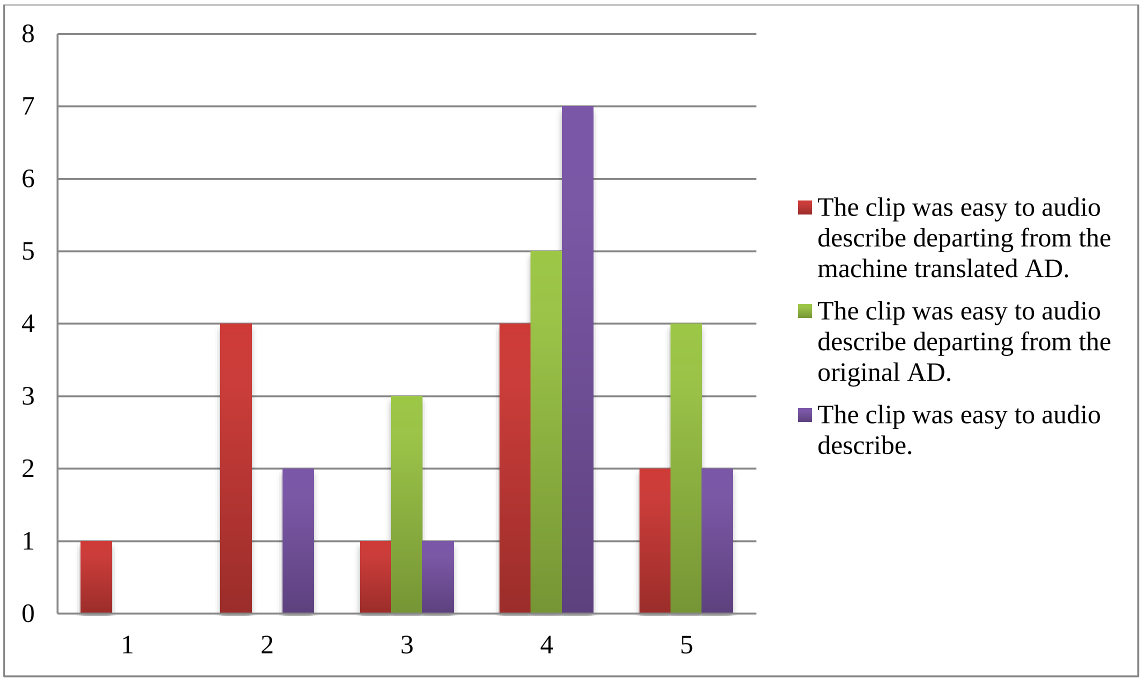

Figure 3 Self-reported ease of audio description in each scenario

The frequency chart indicates a higher variability in the answers for the post-editing task, ranging from 1 to 5 with the same number of participants selecting 2 and 4 (four participants), for instance. Regarding translation, the chosen values range from 3 to 5, showing that participants find it an easy task in a more unified way. Despite participants not having been trained in translating audio descriptions, they have a strong background and translation training, and this possibly affects the results. Finally, regarding AD creation, the vast majority selected 4 (7 out of 12), proving again that this is viewed as an easier task compared with post-editing, despite having only taken one course on audio description at MA level. When mean and median values are considered, the results are the following: AD creation (mean=3.75, median=4), AD translation (mean=4.08, median=4), AD PE (mean=3.17, median=3.5). Two additional statements looked further into the ease of the task: on the one hand, participants were asked their level of agreement with the statement "The source text was easy to translate", obtaining a mean value of 4.08 and a median of 4. When the same formulation was used for post-editing ("The machine-translated text was easy to post-edit"), the values were 2.67 and 2.5, respectively. This proves again how translation is perceived as an easier task than post-editing, at least when referring to the texts provided in the experiment. Needless to say that many factors can impact on these results: on the one hand, previous training of the participants; on the other, the quality of the machine translation output. In this regard, when participants were asked to report their level of agreement with the statement "The machine-translated text required no post-editing" on a 5-point Likert scale, values were almost the lowest possible (mean= 1.08, median=1). In response to the sentence “The machine-translated text was fluent Catalan”, the mean was 1.75 and the median was 2, and figures were slightly higher when assessing the statement “All the information in the source text was present in the machine-translated text” (mean=3.25, median=3.5).

Conclusions

This article has presented an experiment in which the efforts of creating an AD, translating an AD and post-editing a machine-translated AD were compared, with the ultimate aim of exploring whether machine translation could be satisfactorily deployed in audio description. After presenting an overview of the current state of the art, and describing the experimental design, results were discussed.

Post-editing is generally considered to be faster than human translation (Daems et al. 2015: 31), and many existing experiments prove this (de Sousa, Aziz and Specia 2011, Koglin 2015). In our test, despite being the fastest option, the differences are extremely low: on average post-editing takes only four minutes less than translating, and the difference between post-editing and creating an AD is just a few seconds. However, other indicators tend to present wider differences, and both technical and cognitive effort seem to be less demanding in post-editing. Moreover, even though no statistically significant differences were found in most cases – probably due to sample size limitations –, post-editing is usually the task displaying the most homogeneous results and, therefore, less variability, which makes the mean values obtained more reliable. It would therefore seem that implementing machine translation for audio description may be a feasible solution, or at least one which merits further investigation.

Nonetheless, if subjective effort assessments are to be considered, post-editing is generally expected to be the most demanding task in terms of effort, an idea that is reinforced once the task is performed, when the effort perception has the lowest value. This is in sharp contrast with objective data, and makes us think of the need to carry out studies which not only provide numerical data on already established indicators that can be objectively measured but also gather feed-back from users. New technological solutions cannot only be measured in terms of time or productivity, and this also applies to a possible implementation of machine translation in the audio description field.

Due to its exploratory nature, this experiment has several limitations, opening the door to further research. First of all, as already stated above – and despite it being common practice in this kind of research (Temizöz 2012) –, the small sample size does not allow statistically robust conclusions to be drawn. A larger sample would allow for sounder extrapolations to be made.

Secondly, the participants’ profile has undoubtedly had an impact on the results. It was decided to use postgraduate students from the same MA programme in order to ensure a uniformly comparable sample. It remains to be seen what would happen if more experienced translators, post-editors or audio describers were selected for the test rather than AV students. One could hypothesise that time spent on the tasks by professionals compared with novices would be lower, as demonstrated by Moorkens and O’Brien (2015), but, as the same authors point out, professional attitudes towards technology may be more negative. Additionally, it would be interesting to find out whether there would be any differences between professionals with different profiles (audio describers, translators, post-editors), as it would be considerably more difficult to find professionals with completely comparable experience in these three fields.

Thirdly, evaluating the output quality, not just the process, as in this paper, would be a necessary next step. Assessing the output for the three scenarios under analysis, both by experts and by end users – mainly blind and visually impaired audiences –, would undoubtedly offer more information on this topic.

Finally, it would be worthwhile replicating the same experiment with other data sets and language pairs, to get a wider overview of the possibilities of machine translation in this new field. Many research possibilities emerge, but this paper can be considered a first step in a rather under-researched topic in the field of audiovisual translation.

Acknowledgements

We would like to thank Märielle Leijten and Eric Van Horenbeeck for their help and support with the keystroke-logging tool Inputlog. Special thanks are also due to the students of the MA in Audiovisual Translation who volunteered to take part in the experiments. This work was supported by the Spanish Ministerio de Ciencia e Innovación project ALST (reference code FFI2012-31024). Anna Matamala is a TransMedia Catalonia member, funded by Generalitat de Catalunya (2014SGR27).

References:

ARMSTRONG, Stephen, WAY, Andy, CAFFREY, Colm, FLANAGAN, Marian, KENNY, Dorothy, O’HAGAN, Minako. 2006. Improving the quality of automated DVD subtitles via example-based machine translation [online]. In *Proceedings of Translating and the Computer*, 2006, vol. 28 [cit. 2015-12-19], no page numbers. Available at: <http://www.mt-archive.info/Aslib-2006-Armstrong.pdf>.

CARL, Michael, DRAGSTED, Barbara, ELMING, Jakob, HARDT, Daniel, JAKOBSEN, Arnt Lykke. 2011. The Process of Post-Editing: a Pilot Study [online]. In *Proceedings of the 8th International NLPSC Workshop. Copenhagen Studies in Language*. 2011, vol. 41 [cit. 2015-12-19], pp. 131-142. Available at: <http://www.mt-archive.info/NLPCS-2011-Carl-1.pdf>.

CHUKHAREV-HUDILAINEN, Evgeny. 2014. Pauses in spontaneous written communication: A keystroke logging study [online]. In *Journal of Writing Research*. 2014, vol. 6, no. 1 [cit. 2015-12-19], pp. 61-84. Available at: <http://dx.doi.org/10.17239/jowr-2014.06.01.3>.

DAEMS, Joke, VANDEPITTE, Sonia, HARTSUIKER, Robert, MACKEN, Lieve. 2015. The impact of machine translation error types on post-editing effort indicators [online]. In *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)*. 2015 [cit. 2015-12-19], pp. 31-45. Available at: <http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf>.

DAM-JENSEN, Helle, HEINE, Carmen. 2013. Writing and translation process research: Bridging the gap [online]. In *Journal of Writing Research*. 2013, vol. 5, no. 1 [cit. 2015-12-19], pp. 89-101. Available at: <http://dx.doi.org/10.17239/jowr-2013.05.01.4>.

DELGADO, Héctor, MATAMALA, Anna, SERRANO, Javier. 2015. Speaker Diarization and Speech Recognition in the Semi-Automatization of Audio Description: An Exploratory Study on Future Possibilities. In *Cadernos de Traduçao*, vol. 35, no. 2, pp. 308-324.

DEL POZO, Arantza. 2014. *SUMAT final report* [online]. 2014 [cit. 2015-06-01]. Available at: <http://www.sumat-project.eu/uploads/2014/07/D1-5_Final-Report-June-2014.pdf>

DE SOUSA, Sheila C. M., AZIZ, Wilker, SPECIA, Lucia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles [online]. In *Proceedings of Recent Advances in Natural Language Processing*. 2011 [cit. 2015-12-19], pp. 97-103. Available at: <http://aclweb.org/anthology/R11-1014>.

EUROPEAN COMMISSION. n.d. MT@EC A machine translation service, covering all of the EU's official languages [online] [cit. 2015-12-19]. Available at: <http://ec.europa.eu/isa/ready-to-use-solutions/mt-ec_en.htm>.

FERNÁNDEZ-TORNÉ, Anna. Forthcoming. Machine Translation Evaluation through Post-Editing Measures in Audio Description.

FERNÁNDEZ-TORNÉ, Anna, MATAMALA, Anna. 2015. Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan. In *The Journal of Specialised Translation* [online]. 2015, vol. 24 [cit. 2015-12-19], pp. 61-88. Available at: <http://www.jostrans.org/issue24/art_fernandez.pdf>.

GUERBEROF, Ana. 2009. Productivity and quality in MT post-editing [online]. In *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. 2009 [cit. 2015-12-19], no page numbers. Available at: < http://www.mt-archive.info/MTS-2009-Guerberof.pdf>.

GUERRA, Lorena. 2003. *Human translation versus machine translation and full post-editing of raw machine translation output*. MA diss. Dublin : Dublin City University, 2003.

HOUSLEY, Jason K. 2012. Ruqual: A system for assessing post-editing [online]. In *All Theses and Dissertations*. 2012 [cit. 2015-12-19], no. 3106. Available at: <http://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=4105andcontext=etd>.

JANKOWSKA, Anna. 2013. Tłumaczenie skryptów audiodeskrypcji z języka angielskiego jako alternatywna metoda tworzenia skryptów audiodeskrypcji. [Translation of audio description scripts from English as an alternative method of audio description scripts creation]. PhD diss. Cracow : Jagiellonian University, 2013.

KOGLIN, Arlene. 2015. An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors [online]. In *Translation and Interpreting*. 2015, vol. 7, no. 1 [cit. 2015-12-19], pp. 126-141. Available at: <http://www.trans-int.org/index.php/transint/issue/view/29>.

KOPONEN, Maarit. 2015. How to teach machine translation post-editing? Experiences from a post-editing course [online]. In *Proceedings of 4th Workshop on Post-Editing Technology* *and Practice (WPTP4)*. 2015 [cit. 2015-12-19], pp. 2-15. Available at: <http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf>.

KOPONEN, Maarit, RAMOS, Luciana, AZIS, Wilker, SPECIA, Lucia. 2012. Post-Editing Time as a Measure of Cognitive Effort [online]. In *Proceedings of 1st Workshop on Post-Editing Technology and Practice (WPTP1)*. 2012 [cit. 2015-12-19], pp. 11-20. Available at: <http://amta2012.amtaweb.org/AMTA2012Files/start.htm>.

KRINGS, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent : Kent State University Press, 2001.

LACRUZ, Isabel, DENKOWSKI, Michael, LAVIE, Alon. 2014. Cognitive Demand and Cognitive Effort in Post-Editing [online]. In *Proceedings of the 3rd Workshop on Post-Editing Technology and Practice (WPTP3),* 2014 [cit. 2015-12-19], pp. 73-84. Available at: <http://amtaweb.org/AMTA2014Proceedings/AMTA2014Proceedings_PEWorkshop_final.pdf>.

LACRUZ, Isabel, SHREVE, Gregory M., ANGELONE, Erik. 2012. Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. In *Proceedings of 1st Workshop on Post-Editing Technology and Practice (WPTP1)*. 2012 [cit. 2015-12-19], pp. 29-38. Available at: <http://amta2012.amtaweb.org/AMTA2012Files/start.htm>.

LEIJTEN, Mariëlle, VAN WAES, Luuk. 2013. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. In *Written Communication*, vol. 30, no. 3, pp. 358-392.

MASZEROWSKA, Anna, MATAMALA, Anna, ORERO, Pilar. 2014. *Audio Description. New perspectives illustrated.* Amsterdam : John Benjamins, 2014.

MATAMALA, Anna. 2006. La accesibilidad en los medios: aspectos língüístícos y retos de formación. In PÉREZ-AMAT, Ricardo, PÉREZ-UGENA, Álvaro. *Sociedad, integración y televisión en España*. Madrid : Laberinto, 2006, pp. 293-306.

MATAMALA, Anna. 2015. The ALST Project: Technologies for Audiovisual Translation. In *Translating and the Computer*, November 2015, vol. 37, pp. 79-89.

MOORKENS, Joss, O’BRIEN, Sharon. 2015. Post-editing evaluations: trade-offs between novice and profesional participant [online]. In *Proceedings of the 18th annual conference of the European Association for Machine Translation (EAMT 2015)*. 2015 [cit. 2015-12-19], pp. 75-81. Available at: <https://aclweb.org/anthology/W/W15/W15-4910.pdf>.

NICHOLS, Mike. 2004. [Motion picture]. *Closer*. USA : Columbia Pictures, 2004. Ident. No. CDR 37281.

O’BRIEN, Sharon. 2004. Machine translatability and post-editing effort: How do they relate? In *Translating and the Computer,* November 2004, vol. 26, no page numbers.

O’BRIEN, Sharon. 2006. Pauses as indicators of cognitive effort in post-editing machine translating output. In *Across Languages and Cultures*, 2006, vol. 7, no. 1, pp. 1-21.

O’BRIEN, Sharon. 2009. Eye tracking in translation process research: methodological challenges and solutions. In MEES, Inger M., ALVES, Fabio, GOPFERICH, Susanne, (eds.) *Methodology, technology and innovation in translation process research: a tribute to Arnt Lykke Jakobsen. Copenhagen studies in language*. Copenhagen : Samfundslitteratur, 2009, no. 38, pp. 251-266.

O’BRIEN, Sharon. 2010. Introduction to post-editing: Who, what, how and where to next [online]. In *Association for Machine Translation in the Americas AMTA 2010*. 2010 [cit. 2015-12-19], no page numbers. Available at: <http://amta2010.amtaweb.org/AMTA/papers/6-01-ObrienPostEdit.pdf>.

O’BRIEN, Sharon. 2011. Towards predicting post-editing productivity [online]. In *Machine Translation*. 2011, vol. 25 [cit. 2015-12-19], pp. 197-215. Available at: <http://doras.dcu.ie/17154/1/Towards_Predicting_Postediting_Productivity_Final_2.pdf>.

ORTIZ-BOIX, Carla. 2012. *Technologies for audio description: study on the application of machine translation and text-to-speech to the audio description in Spanish*, MA diss. Barcelona : Universitat Autònoma de Barcelona, 2012.

ORTIZ-BOIX, Carla, MATAMALA, Anna. 2015. Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: A Three-Level Approach. [online]. In *Proceedings* *of 4th Workshop on Post-Editing Technology and Practice (WPTP4)*. 2015 [cit. 2015-12-19], pp. 2-15. Available at: <http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf>.

ORTIZ-BOIX, Carla, MATAMALA, Anna. Forthcoming. Post-editing Wildlife Documentary Films: A New Possible Scenario? In *Jostrans*, vol. 26.

PLITT, Mirko, MASSELOT, François. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. In *The Prague Bulletin of Mathematical Linguistics*, vol. 93, pp. 7-16.

REMAEL, Aline, VERCAUTEREN, Gert. 2007. Audio describing the exposition phase of films. Teaching students what to choose. In *TRANS*, 2007, vol. 11, pp. 73-93.

SALWAY, Andrew. 2004. AuDesc system specification and prototypes, TIWO: Television in Words [online]. 2004, vol. 3 [cit. 2015-12-19]. Available at: <http://www.bbrel.co.uk/pdfs/TIWO_Television_in_Words_Deliverable_3.pdf>.

SPECIA, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort [online]. In *Conference of the European Association for Machine Translation*. 2011, vol. 15 [cit. 2015-12-19], pp. 73-80. Available at: <http://www.mt-archive.info/EAMT-2011-Specia.pdf>.

TATSUMI, Midori, ROTURIER, Johann. 2010. Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship? [online]. In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*. 2010 [cit. 2015-12-19], pp. 43-51. Available at: <http://www.mt-archive.info/JEC-2010-Tatsumi.pdf>.

TAUS, CNGL. 2010. *Machine translation postediting guidelines* [online]. 2010 [cit. 2015-12-19]. Available at: <http://taus-website-media.s3.amazonaws.com/images/stories/guidelines/taus-cngl-machine-translation-postediting-guidelines.pdf>.

TEMIZÖZ, Özlem. 2012. Machine translation and postediting. In *The European Society for Translation Studies Research Committee State-of-the-Art Research Report*.

VOLK, Martin. 2009. The automatic translation of film subtitles. A machine translation success story? [online[. In *JLCL*, 2009, vol. 24, no. 3 [cit. 2015-12-19], pp. 113-125. Available at: <http://www.uzh.ch/news/articles/2008/3000/Volk__MT_of_Subtitles.pdf>.

Anna Fernández-Torné

Department of Translation and Interpreting and East Asian Studies, Universitat Autònoma de Barcelona

La Vinya, 11B

08329 Teià (Barcelona)

Spain

e-mail: anna.torne@gmail.com

Anna Matamala

Department of Translation and Interpreting and East Asian Studies, Universitat Autònoma de Barcelona

Despatx K-1002, Campus de la UAB

08193 Bellaterra (Barcelona)

Spain

e-mail: anna.matamala@uab.cat