"Machine translation evaluation through post-editing measures in audio description", *inTRAlinea* Vol. 18.
Permanent URL: http://www.intralinea.org/archive/article/2200

inTRAlinea [ISSN 1827-000X] is the online translation journal of the Department of Interpreting and Translation (DIT) of the University of Bologna, Italy. This printout was generated directly from the online version of this article and can be freely distributed under the following **Creative Commons License**.

Machine translation evaluation through post-editing measures in audio description

By Anna Fernández Torné (Universitat Autònoma de Barcelona, Spain)

Abstract & Keywords

English:

The number of accessible audiovisual products and the pace at which audiovisual content is made accessible need to be increased, reducing costs whenever possible. The implementation of different technologies which are already available in the translation field, specifically machine translation technologies, could help reach this goal in audio description for the blind and partially sighted. Measuring machine translation quality is essential when selecting the most appropriate machine translation engine to be implemented in the audio description field for the English-Catalan language combination. Automatic metrics and human assessments are often used for this purpose in any specific domain and language pair. This article proposes a methodology based on both objective and subjective measures for the evaluation of five different and free online machine translation systems. Their raw machine translation outputs and the post-editing effort that is involved are assessed using eight different scores. Results show that there are clear quality differences among the systems assessed and that one of them is the best rated in six out of the eight evaluation measures used. This engine would therefore yield the best freely machine-translated audio descriptions in Catalan presumably reducing the audio description process turnaround and costs.

Keywords: accessibility, media accessibility, audio description ad, audiovisual translation, machine assisted translation, post-editing, Catalan language

Introduction

Linguistic and sensorial media accessibility has become part of the European Union agenda in recent years. Cultural and linguistic diversity is being promoted and laws have been passed in different EU countries to ensure a minimum number of audiovisual products are being made accessible for people with hearing or visual disabilities (European Union Agency for Fundamental Rights 2014). Therefore, there is an urge to provide subtitled—both for those that are not hearing impaired and for the deaf and hard of hearing—and audio described products.

Creating subtitles and audio descriptions (AD) from scratch is a time consuming task with an economic impact which not all content providers can —or are willing to— undertake. Further, and linked to the pressure to reduce the costs of making accessible an ever-increasing volume of audiovisual content are the demands to meet shorter deadlines. In order to deal with this threefold issue —increasing volumes, lowering prices and shortened timeframes—, the translation of subtitles from a prepared English template (Georgakopoulou 2010) and the translation of AD scripts (Jankowska 2015) have been tried and proved as an efficient solution.

Applying new technologies, such as translation memory (TM) tools and machine translation (MT), has also been proved effective and profitable in many translation areas in which texts are more repetitive and predictable (technical texts, for instance) by increasing productivity and improving terminology consistency (Choudhury and McConnell 2013). However, the use of TM tools is not at all generalised in the domain of audiovisual translation (AVT) (Hanoulle 2015) and the implementation of MT is opposed by many of its main actors, that is audiovisual translators, who argue that machines will never be able to deliver human-like quality and that it would only lead to lower prices, as has happened with the implementation of TM tools (Bowker and Fisher 2010). However, these prejudices seem to be slowly dissipating in view of their clear usefulness and improved quality results, particularly in the subtitling field (Georgakopoulou 2011).

Considering their potential, researchers in AVT have begun to dig into the possibilities of implementing different technologies to try to allow for higher accessibility. Some projects relating to MT and subtitling have been funded, but very little research has yet been carried out regarding AD and the application of MT, in spite of Salway's conclusions (2004: 6) that '[t]he relatively simple nature of the language used in audio description (simple that is say compared to a novel), may mean automatic translation systems fair [*sic*] better than usual'. Thus, my interest as a researcher is to present a first approach to this new-born research area and to examine whether MT can successfully be used in the AD arena in the Catalan context. Therefore, and according to Temizöz's (2012: 1) report on 'empirical studies on machine translation and the postediting of MT output', the novelty of my work lies mainly in the 'type of text' covered (AD).

My ultimate aim is to compare the effort in three different scenarios: when creating an AD (that is, when translating the visuals into words); when translating an already existing AD (in this case, from English into Catalan); and when post-editing a machine-translated AD, again from English into Catalan. However, when post-editing machine-translated ADs, it is obvious that the choice of the MT engine will have a direct impact on the raw MT output and the subsequent post-editing (PE) effort. This is why a pre-test was carried out in order to select the best engine available for my language pair.

This pre-test is what is described in this article, focusing on the methodology adopted and the various subjective and objective measures used. Assessing the quality of the resulting post-edited versions is beyond the scope of this paper.

This article presents first a short review of the existing work related to human translation and MT in the audiovisual fields of subtitling and AD. It then describes the set-up of the experiment, including the participants involved, the test data used, and the MT engines analysed. It also details the MT output evaluation tasks performed by the testers and the PE tool chosen, followed by the actual development of the pre-test. Next it explains the statistical methods used and discusses the results obtained. It finally presents the conclusions and assesses the opportunities for further research in this field.

Machine-translated audio description: related work

Before contemplating the post-editing of machine-translated ADs, a closer look into the controversy around their human translation is needed. In line with some current practices and working processes in the subtitling market (Georgakopoulou 2010), several researchers defend not only the viability of translating AD scripts (Jankowska 2015; Matamala 2006; Salway 2004), but also the necessity (López Vera 2006).

There are also critics to this proposal: Hyks (2005: 8) argues that 'translating and rewording can sometimes take as long if not longer than starting from scratch' and that '[t]he fact that some languages use many more or sometimes fewer words to express an idea, can drastically affect timings'. Rodríguez Posadas and Sánchez Agudo's (2008) opinion is much more categorical. They talk about 'putting a foreign culture before the Spanish (or the Spanish blind people's) culture' (*ibid.*: 8) and about a 'lack of respect for the blind' (*ibid.*: 16), and argue that translating AD scripts would be more expensive since it would involve not only the translator but also a dialogue writer.

Be it as it may, Remael and Vercauteren (2010: 157) maintain that 'AD translation does happen' and claim that it 'will increase in the (near) future, if only because it may be perceived as a cost-cutting factor by international translation companies, film producers and distributors'. In this sense, it must be stated that this is no longer a mere perception. As a result of the research conducted by Jankowska (2015), it has been proved that visually impaired people accept translated ADs and that it is a less time-consuming and cheaper process than creating them from scratch.

As far as MT is concerned, its implementation has been researched in the subtitling domain as a possible solution. Popowich, McFetridge, Turcato and Toole (2000) were pioneers in presenting a rule-based MT system that provided the translation of closed captions from English into Spanish and concluded that the subtitling domain was already appropriate for the development state of MT systems at that time.

Several European projects have since been developed. MUSA (2002-2004) aimed at 'the creation of a multimodal multilingual system that converts audio streams into text transcriptions, generates subtitles from these transcriptions and then translates the subtitles in other languages' (Languages and the Media 2004: 3). Not long afterwards, eTITLE (2003-2005) was launched. It presented a system that combined MT with TM technologies in the subtitling environment in several linguistic combinations, including English to Catalan. SUMAT (2011-2014) offered an on-line service for subtitling by MT. Its final report stated that results were 'quite positive when measuring quality in terms of objective metrics and rating by professional users, with a significant portion of MT output deemed to be of a sufficient quality to reach professional quality standards through minimal to medium PE effort. Productivity measurement also indicated time gains across the board' (Del Pozo 2014: 40). In turn, the EU-BRIDGE project developed the automatic transcription of TV shows to subtitle them and translate the subtitles into multiple languages.

Apart from these EU funded projects, in the academic sphere O'Hagan (2003) aimed at knowing if language technology could be applied to subtitling, for which she tested 'the usability of freely available MT for creating subtitles mainly by non-professional subtitlers' (*ibid.:* 14). The experiment demonstrated that 'a large proportion of the raw MT outputs of the LOTR [*Lord of the Rings*] English subtitles could be usable as a pure aid to non-English speaking viewers under certain circumstances' (*ibid.:* 14–15), implying that there was a clear scope for potential.

The research by O'Hagan inspired a project developed by Armstrong et al. (2006) to test the feasibility of using a trained example-based MT (EBMT) engine to translate subtitles for the German-English language pair in both directions.

Volk (2009), in turn, explored the application of a trained statistical MT system to translate subtitles in Scandinavian languages. The SMT system was trained with a very large parallel corpus of over 5 million subtitles and results indicated that the machine-translated subtitles were of good quality. Moreover, the translation process was proved to be considerably shortened by the use of such a trained MT system.

De Sousa, Aziz and Specia (2011) went one step further: they assessed the effort involved in translating subtitles manually from English into Portuguese compared to post-editing subtitles which had been automatically translated with the help of CAT tools in the same language pair. They used time as the objective measure for PE effort and their experiments showed that post-editing was much faster than translating subtitles *ex novo*.

However, the implementation of MT in the AD domain has not yet been studied in depth. To the best of my knowledge, only the Master's dissertation by Ortiz-Boix (2012) is devoted to AD and the application of MT. The author compared the quality of machine-translated ADs from Catalan into Spanish based on error analysis. Two free online MT engines without any specific training were used. Google Translate was used as an example of a statistical engine, that is based on statistical models generated after analysing bilingual corpora, and Apertium was used as an example of rule-based engine, that is based on linguistic rules regarding the source and the target languages. The results of these preliminary tests showed that Google Translate made far fewer mistakes than Apertium and proved that applying MT to filmic ADs from Catalan into Spanish would be viable provided that a post-editing by a human was performed before voicing the AD.

Ortiz-Boix's study was carried out within the framework of the ALST (Linguistic and sensorial accessibility) project. This project researches the application of three technologies, including speech recognition, machine translation and speech synthesis, to two oral modes of audiovisual translation, that is, voice-over and AD. The ALST project is where my research is situated, focusing on MT applied to the audio description of feature films as an example of sensorial accessibility.

Experimental Set-Up

Various aspects related to the study design are described next.

Participants

The sample construction was based on one single criterion: participants should be professional translators in the English-Catalan language combination. No professional audio describers were sought for two main reasons. Firstly, audio description is an intersemiotic activity, not necessarily involving an interlinguistic translation, hence not all professional audio describers, either in Catalan or in any other language, are necessarily professional translators. Secondly, since the tasks involved assessing the quality of the raw MT output and transforming it into fit-for-purpose translations, participants had to be professional translators in these languages. No real skills in AD—not even synchronising and adjusting AD units was required—were needed here for the purposes of this test, where the main task was the quality assessment of 5 different MT systems. Therefore, participants were not subjected to any additional requirement.

In the end, the sample was made up of five volunteers: 3 women and 2 men[1], who fulfilled the previous requirements. They were all professional and personal contacts of the researcher and were directly invited via phone call. They were native Catalan speakers and their ages ranged from 24 to 45. None of them had worked professionally in the post-editing of machine-translated texts, providing a homogeneous sample in this regard.

Test data

Since the study aimed to analyse the performance of MT in the field of AD, an AD excerpt had to be chosen. In the selection of the audiovisual product several factors were considered. In the first instance, this experiment is part of a wider project in which other technologies, such as text-to-speech (TTS) in the Catalan context, have been tested. Therefore, a film that had already been audio described both in Catalan (for the TTS AD experiments in which the TTS was compared to the human voiced AD) and in English (for the MT tests) was required. Since my intended target audience were adults and no particular film genre was to be favoured, animated children films were disregarded and a dubbed fiction film belonging to a 'miscellaneous' category according to Salway, Tomadaki and Vassiliou's (2004) classification was chosen: *Closer* (Nichols 2004).

A short clip was selected to minimise participants' fatigue and boredom and to limit the experiment duration. An exhaustive analysis of the film, the AD script and the individual AD units was carried out, and a neutral clip in terms of content (having no potentially distracting such as sex scenes and/or offensive content) and with an AD density of 240 words (1,320 characters distributed among 14 different AD units in 3.09 minutes) was chosen (see Table A.1).

MT engines selection

Although MT performs better with engines that 'are trained with domain-specific memories and glossaries, and work on texts that have been pre-edited following controlled language guidelines' (García 2011: 218), the spirit of the project was to propose a solution that could be used as widely as possible. Therefore, it was decided that only free online MT engines would be used.

A thorough search of the available free online MT engines in the required language pair, that is from English into Catalan, was conducted, and the following engines were found**[2]**:

Yandex Translate, by Yandex

Google Translate, by Google

Apertium, by Universitat d'Alacant

Lucy Kwik Translator, by Lucy Software and Services GmbH

Bing Translator, by Microsoft

This selection included statistically based (Bing Translator, Google Translate and Yandex Translate) and rule-based systems (Apertium and Lucy Kwik Translator). However, no hybrid MT system could be provided, which would have meant a full and comprehensive representation of the current MT models.

The systems will be anonymised in the rest of the article by randomly naming them A to E.

Methodology for MT output quality evaluation

Assessing the quality of MT engines' output poses a major challenge since different approaches exist both in the industry and in the research sphere, and there is no consensus as to which are the best practices.

Both human and automatic measures have been proposed. The most frequent human evaluation measures are sentence-level annotations and include: ranking task (Callison-Burch et al. 2012), error classification (Federmann 2012), PE tasks (either selecting the translation output which is easiest to post-edit or post-editing all outputs) (Popovic et al. 2013), quality estimation (also called expected PE effort) (Federmann 2012; Specia

2011), perceived PE effort (De Sousa, Aziz, and Specia 2011), PE time (Specia 2011), adequacy (Chatzitheodorou and Chatzistamatis 2013), and fluency (Koehn and Monz 2006; Koponen 2010).

Automatic metrics include BLEU (Papineni, Roukos, Ward and Zhu 2002), NIST (Doddington 2002), METEOR (Lavie and Agarwal 2007), and TER (Snover, Dorr, Schwartz, Micciulla and Makhoul 2006), among many others, and are deemed to be 'an imperfect substitute for human assessment of translation quality' (Callison-Burch et al. 2012: 11). However, their use is widely spread because they are easier to implement, faster and cheaper than human evaluation.

In this experiment the focus was on human judgements, both objective and subjective, but automatic metrics were calculated to provide additional data. Thus, the evaluation model resulted in eight scores (see Table 1):

	Automatic	Human
Objective	HBLEU HTER	PE time
Subjective		PE necessity PE difficulty MT output adequacy MT output fluency MT output ranking

Table 1. Evaluation model

On the one hand, Human-targeted Translation Edit Rate (HTER), PE time, PE necessity and PE difficulty were all measurements of the PE effort which each raw MT output required to become a fit-for-purpose translation. On the other hand, Human-targeted Bilingual Evaluation Understudy (HBLEU), MT adequacy, MT fluency and MT ranking focused exclusively on the raw MT output itself.

All objective measures were obtained automatically. HBLEU measured the closeness of a MT to its post-edited versions (Del Pozo 2014). Thus, the higher the HBLEU score of a raw MT output, the closer it was to a professional human translation and therefore it was considered to be better. Its metric ranges from 0 to 1.

HTER measured the distance 'between machine translations and their post-edited versions' (Specia 2011: 74). It counted the number of edits performed to the MT text, including substitutions, shifts, insertions and deletions, divided by the number of words in the post-edited text used as reference. Thus, the more edits performed to a raw MT text (that is, the higher the HTER score), the more effort the PE process was supposed to involve. Its metric also ranges from 0 to 1.

The PE time referred to the total time spent in the post-editing of each AD unit. Again, the more time spent in post-editing, the more effort it was supposed to involve.

In relation to the subjective human assessments, and following Graham, Baldwin, Moffat, and Zobel (2013), four of them (that is all but the ranking task) were presented to participants in the form of 5-point Likert scales to be evaluated according to the participant's level of agreement or disagreement with the given statement. Higher scores represented better results since the statements proposed to participants were formulated so that 'strongly agreeing' (5) or 'agreeing' (4) with them were the most positive answers.

PE necessity assessed to which extent the raw MT output needed to be post-edited in order to obtain a fit-for-purpose target text. As shown in Figure 1, the statement presented to participants was: 'The MT text required no post-editing'. This assessment was meant to be the equivalent to the quality estimation judgement (Federmann 2012) or the expected PE effort appraisal, by which the annotator must decide on the acceptability of a raw MT output in its present condition (Specia 2011).

PE difficulty referred to how difficult post-editing the raw MT output had been. The statement presented to participants was: 'The MT text was easy to post-edit'. This score was inspired by the scale for human translation evaluation in De Sousa, Aziz and Specia (2011) and was related to the perceived PE effort, by which the annotator must assess the effort they have put into post-editing a segment.

The adequacy assessment aimed 'to determine the extent to which all of the content of a text is conveyed, regardless of the quality of the language in the candidate translation' (Chatzitheodorou and Chatzistamatis 2013: 87). The statement presented to participants was: 'All the information in the source text was present in the MT text'.

The fluency judgement (Koehn and Monz 2006; Koponen 2010) tried to convey to what extent a translation flowed naturally and was considered genuine in the target language, without taking into account whether the information was correct and complete in relation to the original text. The statement presented to participants was: 'The MT text is fluent Catalan'.

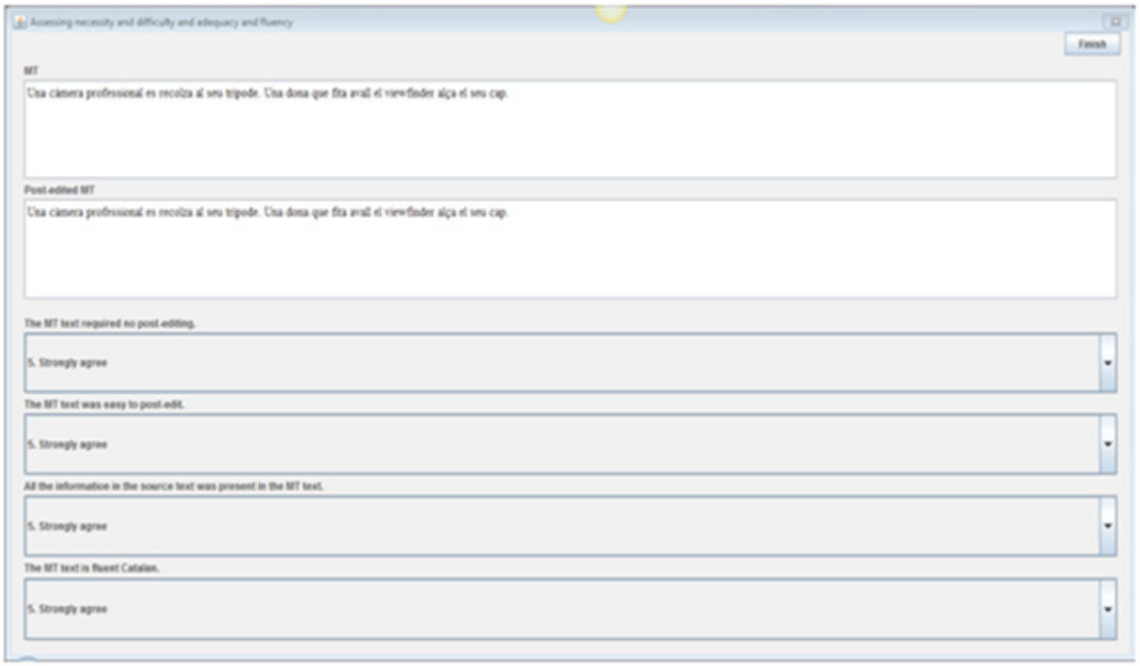


Figure 1. Subjective assessments per AD unit

Finally, the ranking of the raw MT outputs was intended to obtain a classification of each AD unit according to their global quality. Participants were asked to '[r]ank the translation from best to worst, assigning numbers to each unit from 5 (best) to 1 (worst) in the left column', as shown in Figure 2.

	Rank the translation from best to worst, assigning numbers to each unit from 5 (best) to 1 (worst) in the left column.
	A professional camera rests on its tripod. A woman peering down through the viewfinder lifts her head.
1	Unes restes de càmera professionals en el seu tripod. Una dona que mira atentament avall a través del viewfinder ascensors el seu cap.
4	Una càmera professional es basa en el seu trípode. Una dona mirant cap avall a través del visor aixeca el cap.
2	Un professional de la càmera es basa en el seu trípode . Una dona mirant cap avall a través del visor aixeca el cap .
3	Una càmera professional es basa en el trípode. Una dona mirant cap avall a través del visor aixeca el seu cap.
5	Una càmera professional es recolza al seu trípode. Una dona que fita avall el viewfinder alça el seu cap.
	Dan sits stiffly on a stool in front of a screen. The beautiful photographer turns away.
5	Dan rígid asseu en un tamboret davant d'una pantalla. El fotògraf bella allunya.
3	Dan està assegut rígidament en una cadira davant d'una pantalla . El fotògraf bella s'allunya .
1	Dan seu stiffly en un stool davant d'una pantalla. Les voltes de fotògraf boniques fora.
4	Dan seu rígidament en un tamboret davant d'una pantalla. El fotògraf bonic s'allunya.
2	Dan es troba stiffly en un tamboret davant d'una pantalla. La bella fotògraf es converteix distància.

Figure 2. Ranking of raw MT outputs

PE tool selection

Many PE tools, such as Appraise (Federmann 2012), **ACCEPT** (Roturier, Mitchell, and Silva 2013), **TransCenter** (Denkowski and Lavie 2012) and PET (Aziz, De Sousa and Specia 2012), among others, were analysed in order to select the most adequate one for my purposes. Since none of the tools included video and audio options, the AD units could not be accompanied by the real context they were to be inserted in. However, for the aim of this particular test, it was not deemed essential, as no synchronisation or adjustment of the target AD units were asked to the participants.

PET was finally selected for it was a standalone tool and it was absolutely customisable, particularly as far as the assessment questions were concerned. It also allowed for the storage of many other indicators for each AD unit, such as the PE and assessing times, and several edit operations, among others.

Procedure

The experiment was carried out in a real-world environment, which meant that ecological validity was favoured to the detriment of a tighter controlled environment. Participants were informed via email of the tasks to be carried out in a four-hour session.

The test developed as follows. After reading a participant information sheet and signing a consent form approved by the University Ethical Committee, participants were told to download the PE tool and to follow a short training session on the tool. The Catalan dubbed version of the clip with the English AD included in the silent gaps as subtitles was provided for them to watch it. They were next given the script of both the Catalan dialogues and the English AD which they would have to post-edit in written form. They were then told to start the PE tasks. They were allowed to use any resources deemed necessary for the revision (dictionaries, encyclopedias, and so on) and they were instructed not to time-code the AD units. Specific guidelines inspired by the works of O'Brien (2010), De Sousa, Aziz, and Specia (2011), Specia (2011), TAUS and CNGL (2010) and Housley (2012), were also provided:

> Perform the minimum amount of editing necessary to make the AD translation ready for voicing retaining as much raw translation as possible
>
> Aim for a 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 translated correctly.
>
> Basic rules regarding spelling, punctuation and hyphenation apply.

Each participant post-edited five different raw MT versions of the selected AD script excerpt. The order in which MT systems were presented to the participants was balanced to compensate for both the learning effect and the fatigue of the annotators. As indicated above, the excerpt to be post-edited contained 14 AD units, each unit containing one or more sentences**[3]**. After post-editing each unit, participants were asked to provide their evaluations on PE difficulty, PE necessity, MT adequacy and MT fluency, while PE time was automatically calculated by the PE software.

Next, they were asked to rank the translations by assigning numbers to each unit from 5 (best) to 1 (worst). The source English AD unit was displayed, followed by its five different MT versions. The order of the systems was randomised in each unit to prevent any unintentional bias by the participants in relation to a particular system.

Finally, they were asked to fill in a post-questionnaire on participant demographics and subjective opinions. A post-questionnaire was considered more suitable due to the length and complexity of the test.

Statistical methods

Descriptive statistics (mean, median and standard deviation) were computed for the quantitative variables. For the categorical variables —PE necessity, PE difficulty, MT adequacy, MT fluency and MT ranking— percentages were used.

As for the inferential statistics, two different models were applied. On the one hand, a multinomial model with repeated measures was established for each categorical variable with MT system as the explanatory variable. On the other hand, a generalized linear model was established for PE time with MT system as the independent variable.

All results were obtained using SAS, v 9.3 (SAS Institute Inc, USA). For the decisions, significance level was fixed at 0.05.

Results

The best MT system should be the one obtaining the highest HBLEU, the lowest HTER, the lowest PE time, the highest PE necessity, PE difficulty, MT adequacy and MT fluency scores, and the highest position in the ranking. Next, results for each of the items are discussed.

HBLEU

HBLEU metrics were obtained using the Language Studio™ Pro Desktop Tools package, by Asia Online. Table 2 shows that D obtained the highest scores. Therefore, its raw MT output can be considered the best version.

A	B	C	D	E
0.50	0.65	0.60	0.72	0.64

Table 2. HBLEU scores

These scores are deemed to be high. However, as stated by Del Pozo (2014), '[a]s hBLEU scores are measured on post-edited files, they are expected to be higher than the BLEU scores on test sets, as there should be a higher amount of common n-grams in a transformed (that is post-

edited) reference text than in an independently translated reference' (p. 22).

HTER

HTER metrics were obtained using the Language Studio™ Pro Desktop Tools package, by Asia Online. Table 3 shows that D presented the lowest score, which means that its MT outputs were the ones which needed less editing to get to a fit-for-purpose solution.

A	B	C	D	E
0.35	0.25	0.29	0.21	0.26

Table 3. HTER scores

PE time

When comparing the MT engines in terms of the time needed to post-edit their outputs, B produced the translations that required less time to be post-edited, followed by E, D, A and C, that is B obtained the best results (see Figure 3). On average, post-editing an AD unit translated by B took 60 per cent of the time of post-editing one translated by C. C also rendered the highest variability in the PE time (stdev=89.68).

Still, no statistically significant differences were found among the systems. This means that from a statistical point of view no particular MT system could be considered best.

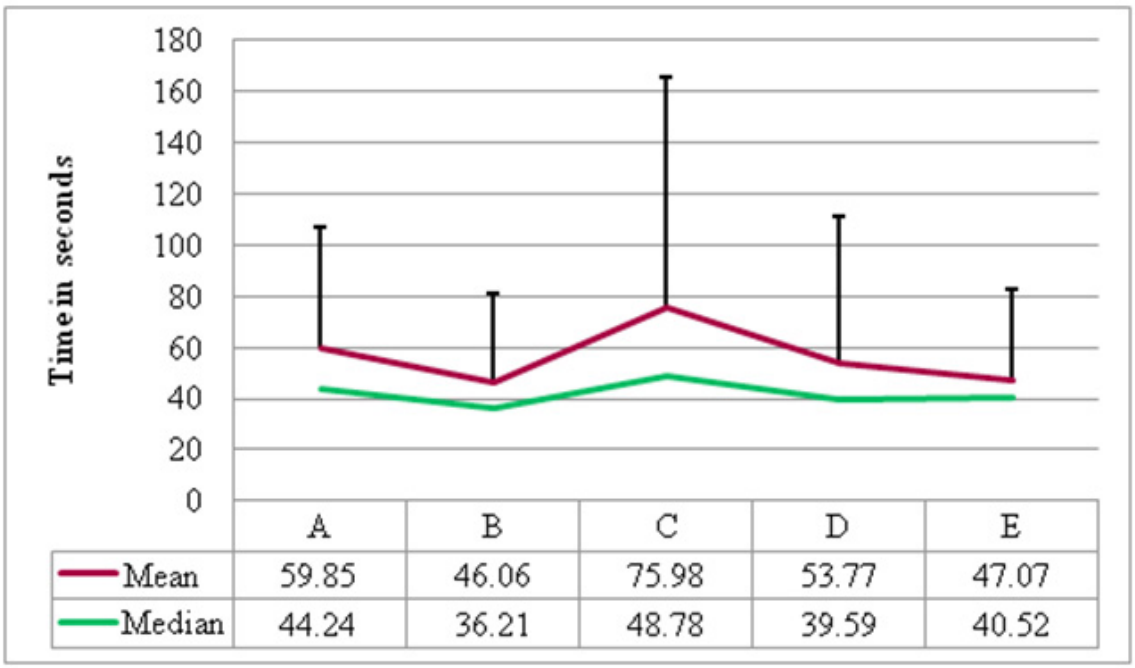


Figure 3. Mean and median PE times per system with standard deviation error bars

PE necessity

According to Figure 4, which shows the frequency of each score for the PE necessity assessment, more than 44 per cent of the AD units translated by D obtained a higher score (scores 4 and 5), that is participants agreed and strongly agreed with the statement 'The MT text required no post-editing' on 32 occasions out of 70. No other system obtained such good results, with E getting higher scores only in 31 per cent of the sentences (22 out of 70), C in 22 per cent of them (16 out of 70), B in 13 per cent of them (9 out of 70), and A in 4 per cent of them (3 out of 70).

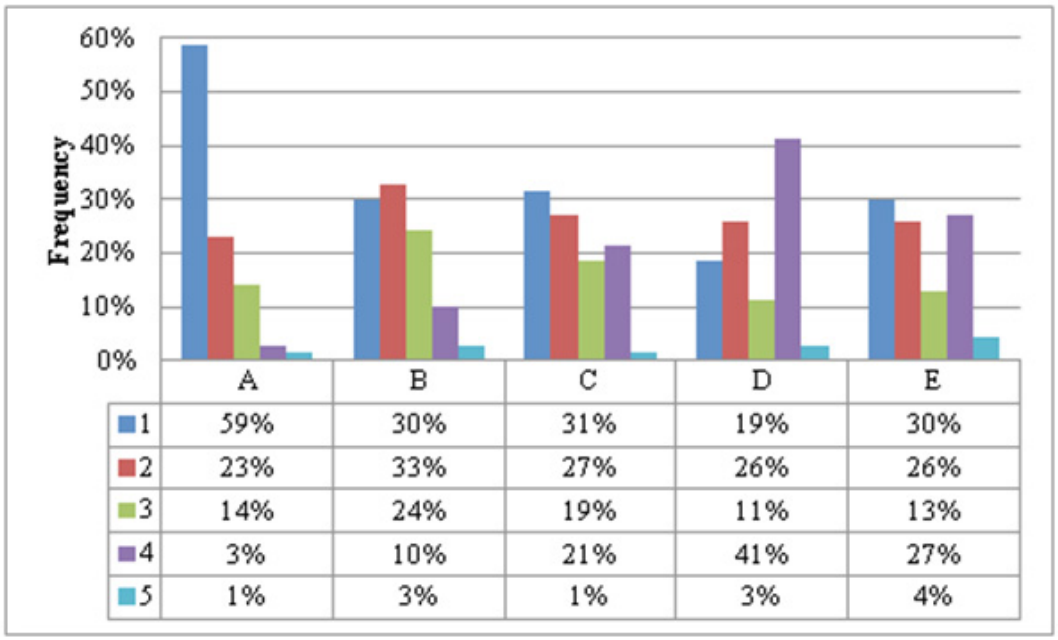


Figure 4. PE necessity scores frequency

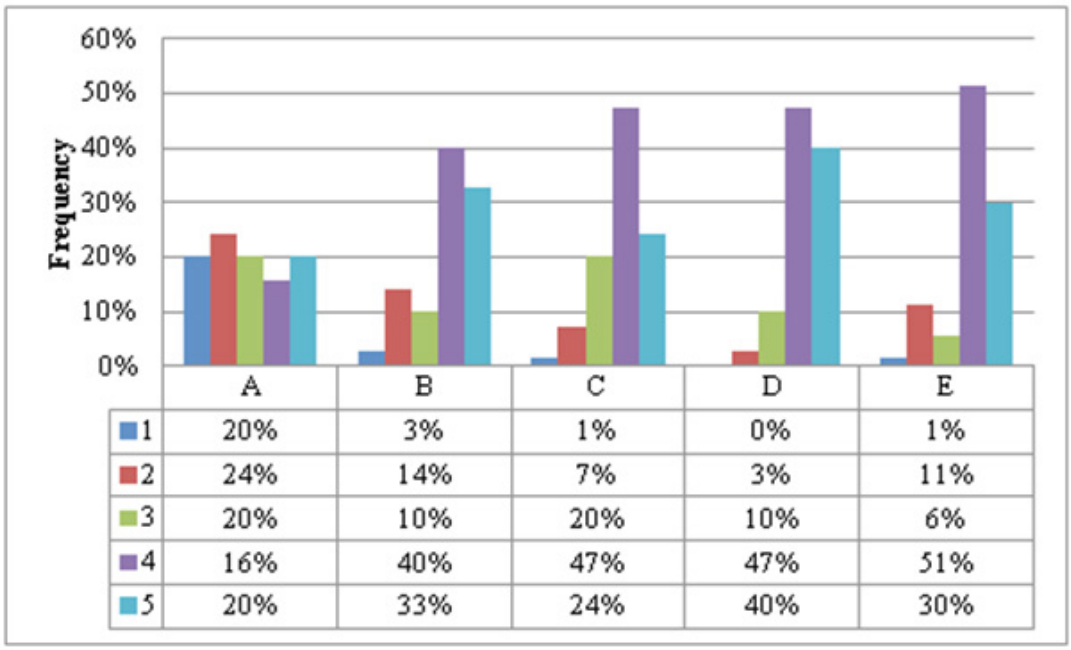
Differences between D and all other MT systems were statistically significant. The odds of obtaining higher scores in D was higher than in any other system, which means that D could be considered the best one as far as PE necessity is concerned (see Table A.2**).**

PE difficulty

PE difficulty scores showed that, although some PE was needed in many occasions, correcting the sentences to get a fit-for-purpose target text was not considered to be a difficult task in most cases.

Figure 5 shows the frequency of each score for the PE difficulty assessment. D obtained the highest frequency for 4 and 5 scores (87 per cent, 61 sentences out of 70), closely followed by E (81 per cent, 57 sentences out of 70), B (73 per cent, 51 out of 70), C (71 per cent, 50 sentences out of 70) and A (36 per cent, 25 out of 70).

In addition, it is worth noticing that no participants assessed D with a 1 score, which means that in no case participants strongly disagreed with the statement 'The MT text was easy to post-edit'. This highlights the fact that none of the sentences translated by D was found very difficult to post-edit by the participants, which did not happen with any other MT engine.



	A	B	C	D	E
1	20%	3%	1%	0%	1%
2	24%	14%	7%	3%	11%
3	20%	10%	20%	10%	6%
4	16%	40%	47%	47%	51%
5	20%	33%	24%	40%	30%

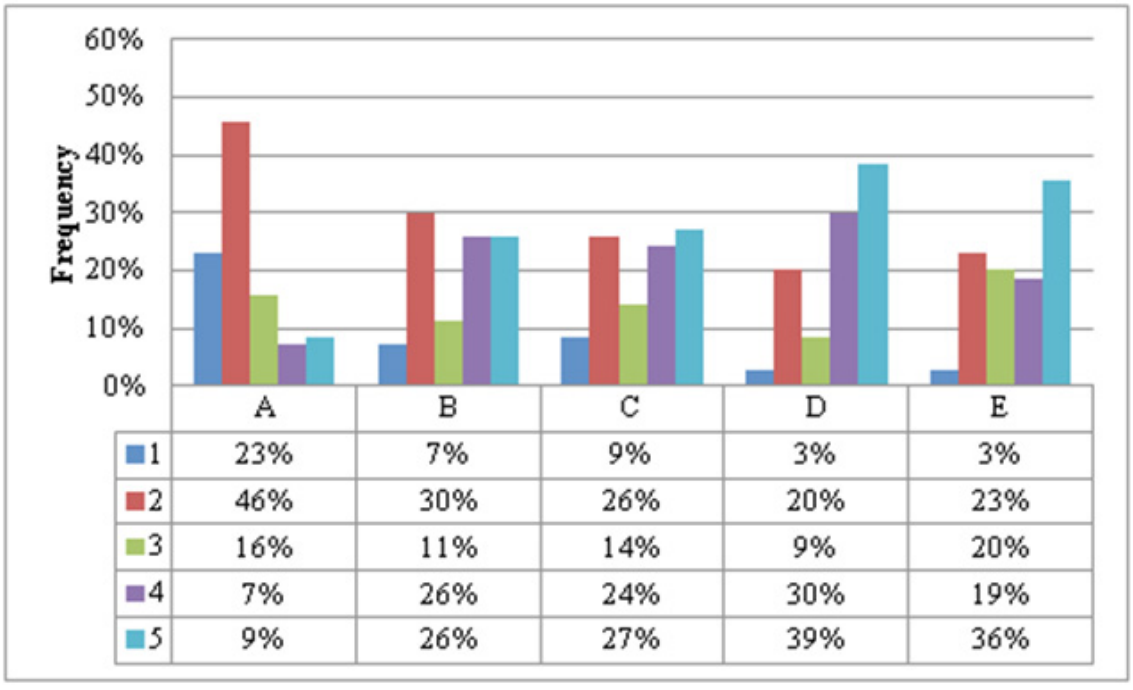
Figure 5. PE difficulty scores frequency

D obtained statistically better scores than B, C, E and A. Thus, D is considered to have the best scores in terms of PE difficulty (see Table A.3).

Adequacy

Taking into account the amount of information of the source actually conveyed in the target text, participants considered that D's MT output presented all or almost all the information of the source AD unit in 69 per cent of the cases (48 out of 70). Figure 6 also shows that D had the highest frequency of 5-score occurrences and the lowest frequency of 2-score occurrences.

Descriptive statistics match with inferential statistics in that D has statistically higher scores than A, B and C, but it is not statistically different from E (see Table A.4).

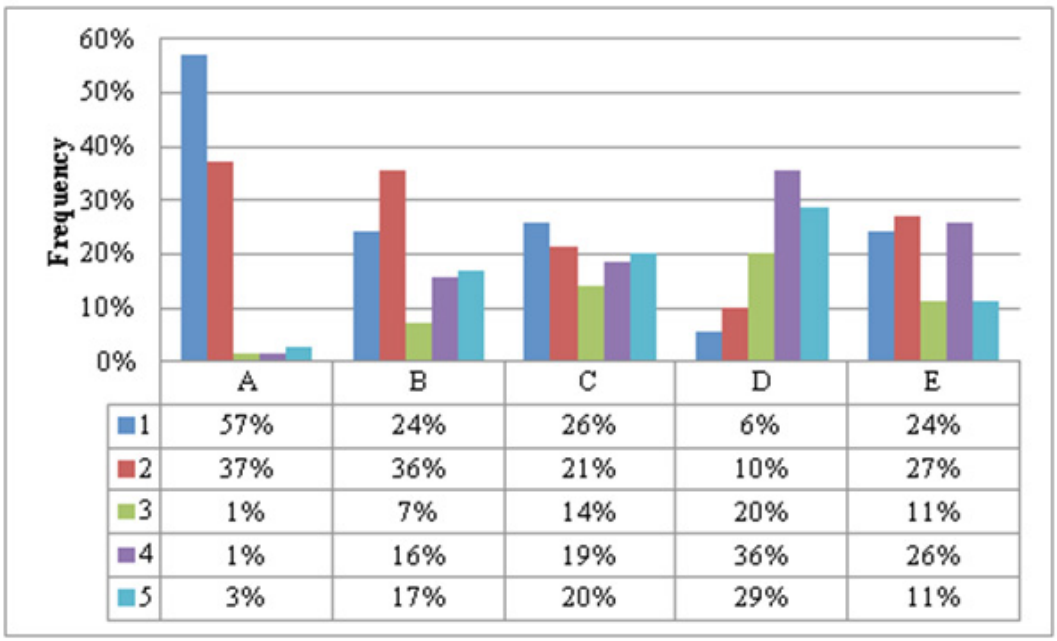


	A	B	C	D	E
1	23%	7%	9%	3%	3%
2	46%	30%	26%	20%	23%
3	16%	11%	14%	9%	20%
4	7%	26%	24%	30%	19%
5	9%	26%	27%	39%	36%

Figure 6. Adequacy scores frequency

Fluency

In terms of fluency, results were quite similar to those of adequacy. Figure 7 shows that D presented the highest frequency of higher scores (65 per cent, 55 out of 70) and the lowest frequency of 1and 2-score occurrences (16 per cent, 11 out of 70), with a total of 20 raw MT outputs being considered fluent Catalan. Inferential statistics, again, confirm these results: D obtained statistically the highest scores (see Table A.5).

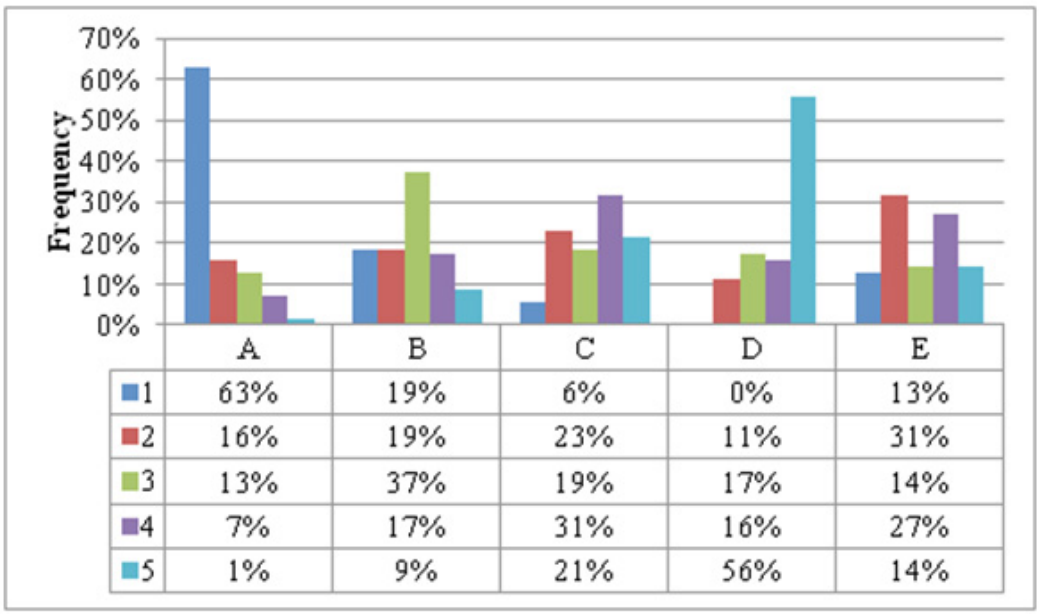


	A	B	C	D	E
1	57%	24%	26%	6%	24%
2	37%	36%	21%	10%	27%
3	1%	7%	14%	20%	11%
4	1%	16%	19%	36%	26%
5	3%	17%	20%	29%	11%

Figure 7. Fluency scores frequency

Ranking

According to Figure 8, 56 per cent of D's raw MT outputs ranked the best ones (39 out of 70), with none of its translations being ranked the worst. These results were statistically confirmed (see Table A.6).



	A	B	C	D	E
1	63%	19%	6%	0%	13%
2	16%	19%	23%	11%	31%
3	13%	37%	19%	17%	14%
4	7%	17%	31%	16%	27%
5	1%	9%	21%	56%	14%

Figure 8. Ranking scores frequency

Conclusions and further research

The aim of the experiment presented in this paper was to propose and implement a methodology which would allow for the selection of the best MT engine to be used in the AD field for the English-Catalan language pair. Participants performed the PE of each MT engine output for the AD script of a 3-minute-long clip and assessed the 14 AD units in terms of PE necessity, PE difficulty, MT adequacy and MT fluency. They also ranked the MT segments from best to worst, and HBLEU, PE time and HTER were automatically computed.

In view of the results exposed above, D was found to be the best MT engine in four out of the five subjective human assessments used in the evaluation (highest PE necessity, PE difficulty and MT fluency scores, and ranking), with the last of the assessments, that is adequacy, presenting higher scores than 3 of the remaining MT systems. In relation to the objective assessments, D also obtained the highest HBLEU scores and outperformed the remaining MT systems in terms of the number of edits needed to get a fit-for-purpose translation. It was just in the PE time score where no statistically significant differences could be found among the MT systems being studied.

However, the study has several limitations, which gives scope for further research and improvement. The first constraint is the number of participants. Increasing it would be desirable to attain a more thorough evaluation, but this was approached as a test previous to the main experiment (Fernández-Torné forthcoming) in which a small sample of five participants was preferred to a subjective decision by the researcher. A second restraint is the test data. Including different AD data sources, such as clips from other film genres, series and documentaries, would also improve the reliability of the test results. In relation to the experimental design, trying to further reduce fatigue in participants by avoiding repetition inasmuch as possible would also be advisable. Other automatic metrics apart from HBLEU and HTER could also be computed for the sake of balancing automatic metrics and human assessments.

Despite these limitations, this article has provided a methodological framework for the evaluation of MT engines in the audiovisual translation field, and more specifically in AD that can be replicated in the future. Needless to say that the study of MT in AVT, and more specifically in AD, is in its infancy, and there are many research possibilities to be explored. For instance, it would be interesting to prove if pre-editing the source texts in the AD field would actually influence on the PE effort, as stated by O'Brien (2010) in other translation domains.

It would also be interesting to see whether the professional profile of the participants, that is having previous professional experience in MT PE or in AD creation, would have an impact on the assessment and final selection of the MT system. As far as the PE instructions are concerned, including the synchronisation (time-coding) and adjustment of the post-edited AD units should also be taken into account, since it is an essential part in AD. In this sense, the development of a PE tool with audiovisual capabilities would actually be much recommended.

Additionally, it would be worth researching the performance of D compared to other MT systems specifically trained with data belonging to the AD domain. As stated by Groves, **'[t]he quality of MT is highly dependent on the quality of the data used for training'** (2011, min. 5.20). Establishing an English-to-Catalan AD corpus would be basic, for which as many English ADs as possible should need to have been previously translated into Catalan. In the absence of such AD translations corpus, the translations of the audiovisual products' scripts could be used to feed the MT systems.

All in all, the test has evaluated the quality of five MT systems by means of automatic metrics and human assessments. Results show that there are clear quality differences among the systems assessed and that D is the best rated in six out of the eight evaluation measures used. This engine would therefore yield the best freely machine-translated ADs in Catalan presumably reducing the AD process turnaround time and costs when compared with the standard process of AD creation. This is what will be researched in our next experiment.

Appendix

AD unit		Duration (seconds)	Words	Characters
1	A professional camera rests on its tripod. A woman peering down through the viewfinder lifts her head.	4.600	17	102
2	Dan sits stiffly on a stool in front of a screen. The beautiful photographer turns away.	4.240	16	88
3	Dressed all in black, Dan puts back his cigarette packet back in his jacket pocket and eyeing the photographer, who is in her thirties, tall and slim, with a chiselled large-featured face. He sits back down. She studies him with a glint in her eye.	11.240	45	248
4	She smiles warmly.	1.040	3	18
5	She nods. Dan stares steadily at her unsmiling. As she turns away again he gets to his feet and crosses the studio.	6.240	22	115
6	Dan looks at some of her photos, which hang on the walls. They are mainly of people.	4.160	17	84
7	Dan wanders back towards the stool and sits.	2.320	8	44
8	She looks coolly at him.	1.480	5	24
9	He straightens his back as she continues to take pictures. She tilts her head to one side regarding him thoughtfully.	5.280	20	117
10	He raises them again flashing a smile. The photographer steps purposefully towards him and adjusts his tie. He looks up at her.	6.600	22	127
11	She goes back to her camera and looks through the viewfinder at Dan. Then lifts her head to look directly at him.	5.040	22	113
12	He stands. She raises the camera on the tripod.	2.200	9	47
13	Dan's piercing eyes dart to one side then fall on the photographer, who meets his gaze and smiles softly, her eyes glistening.	6.680	22	126
14	Her smile gone, she stands motionless, her eyes still fixed on him.	3.160	12	67

Table A.1. Selected clip for the test

Systems	Pr > |t|	OR
D vs A	<.0001	25.3872
D vs B	<.0001	4.6533
D vs C	0.0004	3.4083
D vs E	0.0107	2.3646

Table A.2. PE necessity odds ratio (OR) table

Systems	Pr > |t|	OR
D vs A	<.0001	13.5190
D vs B	0.0118	2.3635
D vs C	0.0082	2.4655
D vs E	0.0627	1.8882
E vs A	<.0001	7.1582
E vs B	0.4958	1.2517
E vs C	0.4182	1.3057

Table A.3. PE difficulty odds ratio (OR) table

Systems	Pr > |t|	OR
D vs A	<.0001	18.7477

D vs B	0.0102	2.3641
D vs C	0.0279	2.0833
D vs E	0.3065	1.4054

Table A.4. MT adequacy odds ratio (OR) table

Systems	Pr > |t|	OR
D vs A	<.0001	22.6552
D vs B	<.0001	4.3459
D vs C	0.0001	3.2819
D vs E	<.0001	3.5360

Table A.5. MT fluency odds ratio (OR) table

Systems	Pr > |t|	OR
D vs A	<.0001	55.3710
D vs B	<.0001	7.9808
D vs C	<.0001	3.5386
D vs E	<.0001	6.2693

Table A.6. Ranking odds ratio (OR) table

References

Armstrong, Stephen, Way, Andy, Caffrey, Colm, Flanagan, Marian, Kenny, Dorothy, and Minako O'Hagan (2006) "Improving the quality of automated DVD subtitles via example-based machine translation" in *Translating and the Computer* 28, London, Aslib: no page numbers.

Aziz, Wilker, De Sousa, Sheila Castilho Monteiro, and Lucia Specia (2012) "PET: a tool for post-editing and assessing machine translation" in *Eighth International Conference on Language Resources and Evaluation,* Nicoletta Calzolari et al. (eds), Istanbul, ELRA: 3982–3987.

Bowker, Lynne, and Des Fisher (2010) "Computer-aided translation" in *Handbook of Translation Studies*, Yves Gambier and Luc van Doorslaer (eds), Amsterdam, John Benjamins: 60-65.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Lucia Specia (2012) "Findings of the 2012 Workshop on Statistical Machine Translation" in *Proceedings of the Workshop on Statistical Machine Translation*, Chris Callison-Burch et al. (eds), Montréal, Association for Computational Linguistics: 10–51.

Chatzitheodorou, Konstantinos, and Stamatis Chatzistamatis (2013) "COSTA MT evaluation tool: An open toolkit for human machine translation evaluation", *The Prague Bulletin of Mathematical Linguistics* 100: 83–89.

Choudhury, Rahzeb, and Brian McConnell (2013) *Translation technology landscape report*, De Rijp, TAUS BV.

De Sousa, Sheila Castilho Monteiro, Aziz, Wilker, and Lucia Specia (2011) "Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles" in *Proceedings of Recent Advances in Natural Language Processing*, Galia Angelova et al. (eds), Hissar, RANLP: 97–103.

Del Pozo, Arantza (2014) *SUMAT final report*, Donostia, Vicomtech-IK4.

Denkowski, Michael, and Alon Lavie (2012) "TransCenter: Web-based translation research suite" in *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*, Sharon O'Brien, Michel Simard, and Lucia Specia (eds), San Diego, AMTA: no page numbers.

Doddington, George (2002) "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics" in *Proceedings of Human Language Technology Research*, no editors, San Francisco, Morgan Kaufmann Publishers Inc.: 138–145.

European Union Agency for Fundamental Rights (2014) *Accessibility standards for audio-visual media: Indicators on political participation of persons with disabilities*, Vienna, FRA.

Federmann, Christian (2012) "Appraise: An open-source toolkit for manual evaluation of MT output", *The Prague Bulletin of Mathematical Linguistics* 98: 25–35.

García, Ignacio (2011) "Translating by post-editing: is it the way forward?", *Machine Translation* 25: 217–237.

Georgakopoulou, Yota (2010) "Challenges for the audiovisual industry in the digital age: Accessibility and multilingualism" in *Proceedings of META Forum 2010*, no editors, Brussels, META-Net: no page numbers.

---- (2011) "Challenges for the audiovisual industry in the digital age: The ever-changing needs of subtitle production", *JoSTrans*, Vol. 17, URL: **http://www.jostrans.org/issue17/art_georgakopoulou.php** (accessed 22 June 2015)

Graham, Yvette, Baldwin, Timothy, Moffat, Alistair, and Justin Zobel (2013) "Continuous measurement scales in human evaluation of machine translation" in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, no editors, Sofia, Association for Computational Linguistics: 33–41.

Groves, Declan (2011) *MT at the CNGL* [Video], Santa Clara, TAUS.

Hanoulle, Sabien, Hoste, Véronique, and Aline Remael (2015) "The efficacy of terminology-extraction systems for the translation of documentaries", *Perspectives: Studies in Translatology* 23: no page numbers.

Housley, Jason K. (2012) *Ruqual: A system for assessing post-editing*, PhD diss., Brigham Young University, USA.

Hyks, Veronica (2005) "Audio description and translation: Two related but different skills", *Translating Today Magazine* 4, no. 1: 6–8.

Jankowska, Anna (2015) *Translating audio description scripts: Translation as a new strategy of creating audio description*. Frankfurt am Main, Berlin, Bern, Brussels, New York, Oxford, Peter Lang.

Koehn, Philipp, and Christoph Monz (2006) "Manual and automatic evaluation of machine translation between European languages" in *Proceedings of the Workshop on Statistical Machine Translation,* Philipp Koehn and Christof Monz (eds), New York City, Association for Computational Linguistics: 102–121.

Koponen, Maarit (2010) "Assessing machine translation quality with error analysis", *MikaEL: Electronic proceedings of the KäTu symposium on translation and interpreting* studies, Vol. 4, URL: **https://sktl-fi.directo.fi/@Bin/40701/Koponen_MikaEL2010.pdf** (accessed 22 June 2015)

Lavie, Alon, and Abhaya Agarwal (2007) "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments" in *Proceedings of the Workshop on Statistical Machine Translation*, Chris Callison-Burch et al. (eds), Prague, Association for Computational Linguistics: 228–231.

López Vera, Juan Francisco (2006) "Translating Audio description scripts: The way forward? Tentative first stage project results" in *MuTra 2006 – Audiovisual Translation Scenarios: Conference Proceedings*, Mary Carroll, Heidrun Gerzymisch-Arbogast, and Sandra Nauert (eds),

Copenhagen, MuTra: no page numbers.

Matamala, Anna (2006) "La accesibilidad en los medios aspectos lingüísticos y retos de formación" in *Sociedad, integración y televisión en España,* Ricardo Pérez-Amat and Álvaro Pérez-Ugena (eds), Madrid, Laberinto: 293–306.

Languages and the Media (2004*) New markets, new tools. Post-conference report*, Berlin, Languages and the Media.

Nichols, Mike (2004) *Closer*, USA, Sony Pictures.

O'Brien, Sharon (2010) "Introduction to post-editing: Who, what, how and where to next" in *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, Alon Lavie et al. (eds), Denver, AMTA: no page numbers.

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

O'Hagan, Minako (2003) "Can language technology respond to the subtitler's dilemma? A preliminary study" in *Proceedings of Translating and the Computer* 25 London, Aslib: no page numbers.

Ortiz-Boix, Carla (2012) *Technologies for audio description: study on the application of machine translation and text-to-speech to the audiodescprition in Spanish*, MA diss., Universitat Autònoma de Barcelona, Spain.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Wei-Jing Zhu (2002) "BLEU: a method for automatic evaluation of machine translation" in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, no editors, Philadelphia, Association for Computational Linguistics: 311–318.

Popovic, Maja, Avramidis, Eleftherios, Burchardt, Aljoscha, Hunsicker, SSabine. Schmeier, Sven, Tscherwinka, Cindy, Vilar, David, and Hans Uszkoreit (2013) "Learning from human judgments of machine translation output" in *Proceedings of the Machine Translation Summit XIV, Khalil* Sima'an, Mikel L. Forcada, Daniel Grasmick, Heidi Depraetere, and Andy Way *(eds), Nice, AMTA: 231-238*.

Popowich, Fred, McFetridge, Paul, Turcato, Davide, and Janine Toole (2000) "Machine translation of closed captions", *Machine Translation* 15, no. 4: 311–341.

Remael, Aline, and Gert Vercauteren (2010) "The translation of recorded audio description from English into Dutch", *Perspectives: Studies in Translatology* 18, no. 3: 155–171.

Rodríguez Posadas, Gala, and Carmen Sánchez Agudo (2007) "Traducción de guiones audiodescriptivos: doble traducción, doble traición" in *AMADIS '07 Congress of the Centro Español de Subtitulado y Audiodescripción (CESyA)*, no editors, Granada, CESyA: no page numbers.

Roturier, Johann, Mitchell, Linda, and David Silva (2013) "The ACCEPT post-editing environment: a flexible and customisable online tool to perform and analyse machine translation post-editing" in *Proceedings of the Machine Translation Summit XIV Workshop on Post-editing Technology and Practice*, Sharon O'Brien, Michel Simard, and Lucia Specia (eds), Nice, AMTA: 119-128.

Salway, Andrew (2004) "AuDesc system specification and prototypes", *TIWO: Television in Words)*, Guildford, University of Surrey.

Salway, Andrew, Tomadaki, Elia, and Andrew Vassiliou (2004) "Building and analysing a corpus of audio description scripts", *TIWO: Television in Word*s, Guildford, University of Surrey.

Snover, Mathew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea, and John Makhoul (2006) "A study of translation edit rate with targeted human annotation" in *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas,* Laurie Gerber et al. (eds), Cambridge, AMTA: 223–231.

Specia, Lucia (2011) "Exploiting objective annotations for measuring translation post-editing effort" in *Proceedings of the 15th Conference of the European Association for Machine Translation*, Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste (eds), Leuven, EAMT: 73–80.

TAUS and CNGL (2010) *Machine translation postediting guidelines*, De Rijp, TAUS.

Temizöz, Özlem (2012) *Machine translation and postediting,* Herentals, European Society for Translation Studies.

Volk, Martin (2009) "The automatic translation of film subtitles. A machine translation success story?", *JLCL* 24, no. 3: 113–125.

Notes

[1] Although five evaluators may seem a low number, it is in line with current research in MT (Specia 2011; O'Brien 2011).

[2] Search performed in September 2013.

[3] The analysis was decided to be at the AD-unit level since this is how an AD is divided semantically. Participants could therefore combine several sentences included in one source AD unit or split one sentence of the source AD unit into several target sentences when post-editing according to their needs.

"Machine translation evaluation through post-editing measures in audio description", *inTRAlinea* Vol. 18.

Permanent URL: http://www.intralinea.org/archive/article/2200