Matamala, Anna. 2015. “The ALST project: technologies for audiovisual translation”. In *Proceedings of the 37th Conference Translating and the Computer*, pages 79-89, London, UK, November 26-27, 2015.

Available from:
http://www.tradulex.com/en/pages/Editions-Tradulex-en

http://asling.org

The ALST Project: Technologies for Audiovisual Translation

Anna Matamala
Departament de Traducció i
d'Interpretació i d'Estudis de l'Àsia
Oriental
Universitat Autònoma de Barcelona
anna.matamala@uab.cat

Abstract

This paper presents an overview of the ALST project, in which speech technologies (speech recognition and speech synthesis) and machine translation were implemented in the voice-over of non-fictional genres and in the audio description of films. The paper presents the project rationale, a brief description of the experiments carried out within the project, as well as its main findings.

1 Introduction

Technologies are very often seen as an indispensable aid to the technical translator's work. However, in the field of audiovisual translation (AVT), the inclusion of technologies in the translation workflow is more recent and has not been always welcomed by professionals. This paper presents the rationale and main findings of a small-scale national project (*Accessibilidad Lingüística y Sensorial: Tecnologías para la audiodescripción y las voces superspuestas*, ALST, i.e. Linguistic and Sensorial Accesibility: Technologies for audio description and voice-over) that, with very limited funding (14,040 Euros for a three-year period, 2013-2015), has researched whether certain technologies could positively impact the creation of accessible audiovisual content. "Accessible" is understood here in a broad sense (Orero and Matamala, 2007), including both access for those who do not understand the original language (linguistic accessibility) and access for those who cannot hear or see the audio or video content (sensorial accessibility), be it because of a disability, impairment or a contextual situation.

The selected technologies were speech recognition, speech synthesis, and machine translation, as they were considered to be mature enough for testing. A future scenario was envisaged in which these three technologies could be concatenated in a working flow, and an original input could be semi-automatically transcribed, machine translated and voiced by a text-to-speech system, always with a human revision process after each step.

The selected audiovisual translation modalities were voice-over and off-screen dubbing, and audio description. Voice-over and off-screen dubbing were chosen as instances of audiovisual modalities catering for linguistic accessibility. Voice-over (Franco *et al*., 2010) is a transfer mode used in many countries to revoice non-fictional genres, although Eastern European countries also use it for fictional content. Díaz-Cintas and Orero (2006: 473) define it as a technique "in which a voice offering a translation in a given target language is heard simultaneously on top of the [source language] SL voice". The sound of the original program is reduced to a low level, and it is "common practice to allow the viewer to hear the original speech in the foreign language at the onset of the speech". Voice-over very often coexists in fictional genres with off-screen dubbing, in which the off-screen voice of the narration or

Proceedings of the 37th Conference Translating and the Computer, pages 79–89,
London, UK, November 26-27, 2015.

commentary in the original content is totally deleted and substituted by a target language version (Franco *et al*., 2010). On the other hand, audio description was chosen as an instance of a modality catering for sensorial accessibility. Audio description (AD) consists in rendering into words the visuals of an audiovisual content (Maszerowska *et al.*, 2015). This description or narration of what is seen on screen is included in the silent gaps in the soundtrack, so that users who do not have access to the visuals can understand and enjoy the audiovisual content. The selected modalities share the characteristic that very often they are delivered orally by a narrator or describer who reads a previously prepared script.

The choice of these modalities allowed us to go beyond existing projects in the field of AVT automatisation, which have mainly focused on machine translation of written outputs such as subtitles (Volk, 2008; De Sousa *et al*., 2011; Del Pozo, 2013). In speech synthesis, experiments on audio description have already been carried out (Szarkowska, 2011; Walczak and Szarkowska, 2012), whilst in speech recognition no specific tests within this field have been developed to the best of our knowledge. It is worth stressing out that the *Strategic Research Agenda for Multilingual Europe* (Rehm and Uszkoreit, 2012: 38) explicitly mentions "automatic voice-over" as a research issue worth exploring, and states that in "2020 we will see wide use of automatic subtitling and first successful examples of automatic voice over for a few languages".

An additional characteristic of the project, which is exploratory in nature, is that no specific tools were developed or improved, but existing resources, very often freely available on the Internet, were chosen. Also, special emphasis on the translator or describer and on the end user was made.

Following the structure of the project, the paper is divided in two parts: Section 2 deals with technologies for linguistic accessibility, whilst Section 3 looks deeper into technologies for sensorial accessibility. Each part describes the specific aims and testing carried out for each technology in each modality. Although the project began with a common aim in mind, and both parts ran in parallel, experiments have not been reproduced identically and specificities have emerged during the project development. It must also be acknowledged that many of the experiments have already been described in published or forthcoming papers, where a more detailed analysis can be found. Hence, the value of this contribution is to offer a broad and unified perspective of the project, despite not being so thorough. It is also worth stressing that all experiments have followed procedures approved by UAB's ethics committee.

2 Technologies for Linguistic Accessibility: Voice-over and Off-screen Dubbing

In the field of voice-over and off-screen dubbings, tests with non-fictional genres from English into Spanish were planned, with the following specific aims in mind:

(a) to investigate whether speech recognition, either automatically or via respeaking, could be used to automatically transcribe non-fictional content,
(b) to research whether machine translation could be useful in the translation process, by comparing the effort involved in translation and in post-editing, and by analysing the output quality in both situations, and
(c) to research how end users would receive a documentary revoiced using text-to-speech compared to human voices, as it is standard practice.

2.1 Speech Recognition in Transcribing Non-fictional Genres

This exploratory research aimed to investigate the inclusion of speech recognition in the transcription of non-fictional content, either automatically or via respeaking (Daniluk *et al.*, 2015). Respeaking is defined as "a technique in which a respeaker listens to the original

sound of a live programme or event and respeaks it, including punctuation and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay" (Romero-Fresco, 2011: 1). However, in our project we aimed to apply it to transcribe recorded content, similar to what in the USA is called voice-writing (Sohn, 2004).

An experiment was designed to compare three situations: manual transcription, respeaking, and revision (or post-editing) of a script generated by an automatic speech recognition (ASR) system. A pilot test with five participants allowed improvement of the experiment design. English was the chosen language.

Ten professional transcribers (4 male, 6 females) with no previous experience of respeaking or ASR post-editing took part in the experiment. Two participants' quantitative data and one participant's qualitative data could not be used for technical reasons.

A video interview lasting 12 minutes was split into three four-minute equivalent excerpts. The video included colloquial spoken language and featured two female American hip-hop artists from California talking about their recent work. It was chosen as it reproduces a real-life situation for which no script is available and a transcription for non-fictional content is needed. An automatic transcript was generated using a state-of-the-art SR system that had not been trained specifically for this content. Although this was expected to affect the results negatively, it was done on purpose as to see how an existing system would perform. Dragon Naturally Speaking 12 Premium was used to respeak.

Participants were received in a computer lab in London and were handed a short pre-questionnaire on demographic information. They were provided with a 30-minute training session on respeaking and then they were requested to fill in a pre-questionnaire that gathered subjective opinions on the three methods involved in the test. They were then instructed to transcribe three excerpts using the three methods (manual transcription/respeaking/ASR post-editing), with the order of tasks and videos being randomized and balanced across participants. Time spent on each task was controlled, and a maximum of 30 minutes was established for each task. At the end of the test a post-questionnaire was distributed to gather additional subjective opinions. Data gathered included: time spent on each task, and ratio "minutes spent on the transcription per minute of original content", as well as qualitative data on users' opinions.

Results indicate that manual transcription was the fastest option (7'39'' spent on transcribing one minute of original content), followed by respeaking (8'36'') and ASR post-editing (9'36''). It is worth highlighting that respeaking is not far from manual transcription, and it was also the method that allowed more participants to complete the task.

Regarding subjective data, it is interesting to observe the participants' replies to a set of identical questions before and after the task (see Table 1).

Results indicate that transcribers perceive current practices (manual transcription) as too time consuming, and are willing to embrace other methods. Respeaking is perceived as a useful tool to transcribe documentaries, both before and after the task, although mean values drop slightly. ASR is also considered useful but the drop after the task is higher, probably due to the testing conditions.

Apart form the previous questions, participants were specifically asked on a 5-point Likert scale about their perceptions in terms of effort involved and boredom, as well as accuracy and overall quality of the transcripts they had generated. Respeaking got the best scores in perceived effort (2.89) and boredom (2.22), whilst manual transcription scored higher in accuracy (4.22) and overall quality (4.33). An in-depth analysis of the results is provided by Matamala *et al.* (forthcoming), who highlight the need for further research in this field.

Statement	Pre-task	Post-task
Manual transcribing is too time consuming	3.4	3.2
Respeaking could be a useful tool to transcribe documentaries	4.5	3.8
Automatic speech recognition could be a useful tool to transcribe documentaries.	4.1	2.7
Respeaking could speed up the process of transcription	4.5	3.9
Automatic speech recognition could speed up the process of transcription	4.1	2.1
Respeaking could increase the accuracy of transcriptions	3.8	2.9
Automatic speech recognition could increase the accuracy of transcriptions	3.0	2.2
Respeaking could increase the overall quality of transcriptions	3.4	3.1
Automatic speech recognition could increase the overall quality of transcriptions.	2.8	2.5

Table 1. Pre-task and post-task opinions (mean values on a 5-point scale, 5 being "completely agree with the statement")

2.2 Machine Translation in Wildlife Documentaries (Voice-over and Off-screen Dubbing)

This experiment was divided in two phases. The first phase compared the effort involved in translating versus post-editing wildlife documentaries excerpts from English into Spanish. Wildlife documentaries were selected after a preliminary study by Ortiz-Boix (forthcoming) proved the feasibility of applying machine translation to this genre.

Following Kring's (2011) proposal on how to measure post-editing effort, effort was considered to include temporal effort (time spent on each task), technical effort (keystroke, mouse movements and clicks for each task), and cognitive effort (pause to word ratio, and average pause ratio, according to Lacruz *et al*., 2014a, 2014b).

Twelve MA students (6 male, 6 female) specialising in AVT participated in the study. They had all taken a course on voice-over in which they had been trained to translate wildlife documentaries. Two 2-minute equivalent excerpts from the documentary *Must Watch: A lioness adopts a baby antelope* were used. Both excerpts were machine translated from English into Spanish by Google Translate as, according to a pre-test (Ortiz-Boix, forthcoming), it was the best free online available MT engine for this language pair and genre at the time the experiment took place. Keyboard logging data were gathered using Inputlog (Leijten and Van Waes, 2013).

Participants were received in a lab simulating real-life working conditions. They were required to translate an excerpt and post-edit another one using a text processor template, balancing the order of presentation and clips across participants. Specific instructions on the output format as well as post-editing/translation guidelines were given. Twenty valid Inputlog files were collected.

Data were analysed independently for each excerpt and globally (considering both excerpts). Results show that post-editing is faster (1,964.525 seconds for post-editing vs. 2,178.116 seconds for translation), although results are only significant in the first excerpt. For both technical and cognitive effort, post-editing requires less effort: 4,025.784 mouse clicks, movements and keystrokes for translation vs. 2,706.565 mouse clicks, movements and keystrokes for post editing (technical effort); 2.756 points between pause to word ratio and

average pause ratio for translation vs. 1.583 points between pause to word ratio and average pause ratio for translation. However, differences are only statistically significant for the first excerpt (not for the second one), and when taking into account all the data. An in-depth analysis per type of effort and per clip is provided in Ortiz-Boix and Matamala (forthcoming a).

The second stage aimed to assess the quality of the output generated in both scenarios. In other words, even if the post-editing effort seems to be lower than translation effort, our aim was to evaluate whether the output quality can be considered comparable. A three-level approach was taken, as explained in Ortiz-Boix and Matamala (forthcoming b): quality assessment by experts, by the dubbing studio, and by end users.

Participants in the first level were six lecturers on MA programmes in AVT at Spanish universities who are also professional translators specialised in the genre. 12 translation and 12 post-editings of two wildlife documentary excerpts (six translation and six post-editings of excerpt one and the same number of excerpt two) were given to the raters. Three evaluation rounds were prepared: in round 1, raters were instructed to read each document and grade it according to their first impression on a 7-point Likert scale. In round 2, raters were asked to correct the documents following a pre-established evaluation matrix based on the MQM error typology (Lommel *et al.*, 2013). After this, they were requested to grade the texts again on a 7-point Likert scale and reply to a questionnaire. In round 3, a final mark between 0 and 10, following Spain's traditional marking system, was requested. A final task consisted in guessing whether the assessed document was a translation or a post-editing, since the nature of the document was blinded.

Results, discussed in detail in Ortiz-Boix and Matamala (forthcoming b), show that, although the quality of both translation and post-editings is considered rather low by experts, no significant differences between post-editings and translations are found. Concerning round 1, while 62.5% of translations are evaluated from "pass" to "excellent", only 51.39% of post-editings are evaluated within this range. However, in round 2, the difference is narrower (56.94% translations vs. 52.78% post-editings). In all instances the median grade for both rounds is a "pass". In the correction carried out at this stage, translation presents a lower number of corrections (mean: 12.861 per document) than post-editings (mean: 17.957). In round 3, the difference in the mark given is again very small: 5.44 for translation versus 5.35 for post-editing. Finally, regarding the post-editing/translation identification task, it is observed that it is easier to identify which texts are translations (58.33% correctly identified) than post-editings (30.55%). The previous data compel us to state that no significant differences are found in both conditions.

In the second-level assessment, the best-rated scripts and videos for each excerpt were sent to a dubbing studio and a professional recording was made. The number of changes made during the recording session was noted down by the researcher, who also took observational notes. Results show that a similar number of changes were made in the first excerpt (6 changes in the post-editing, 5 in the translation). In the second excerpt four changes were made in the translated version. As for the post-editing, the dubbing director considered the synchronisation to be of very low quality and suggested that a re-translation would be needed. Since this was not possible, it was decided to record the excerpt as it was and test whether a negative reaction from audiences would be found in the third level. Therefore, although no quantitative differences are observed between translations and post-editings, data show that translation, at least in the second excerpt, is qualitatively better than post-editing.

In the third-level evaluation, 56 users (28 male, 28 female) were involved. In the data analysis, they were divided into two age groups (group A: <40, group B:>40) because differences in terms of viewing habits and preferences for voice-over were observed in the pre-questionnaire. They watched one post-edited and one translated documentary excerpt, in a

randomized order, without knowing which one they were watching. A questionnaire was distributed after each viewing to test comprehension and enjoyment. Results show that, regardless of the excerpt, version, and age group, users were engaged with the content. Overall findings indicate slightly better results for the translation in terms of enjoyment ("strongly agree" with the statement "I have enjoyed watching the excerpt" in the translated version versus "moderately agree" for the post-editing) and interest (the translated version was considered "very interesting", whilst the post-edited one was considered "pretty interesting"). However, different trends are observed when analysing the data independently for excerpts and age groups (see Ortiz-Boix and Matamala, forthcoming c). When asked which version they prefer, 44.64% of the participants selected the translation, whilst 42.86% selected the post-editing In terms of comprehension, translation also performs slightly better but again different trends emerge in a more specific analysis.

2.3 Text-to-speech in Voicing Documentaries

Tests are currently performed for text-to-speech in documentaries. Participants are asked to assess both natural and artificial voices in terms of overall impression, naturalness, intelligibility, intonation, pronunciation, speech pauses, listening effort, and acceptance. Perceived comprehension and user engagement are also evaluated. A difference is made between excerpts with voice-over (a voice on top of another voice) and off-screen dubbing (an off-screen narrator in which the original English version is not heard). No findings are available at the time of writing this paper.

3 Technologies for Sensorial Accessibility: Audio Description

In the area of AD, the languages involved were English as a source language and Catalan as the target language. The specific aims were the following:

(a) to investigate whether speech recognition could be used to automatically transcribe the AD units, when a script is not available, and propose a new process;
(b) to research whether machine translation could be used, by comparing the effort (and perceived effort) of describers in three scenarios: when creating an AD *ex novo*, when post-editing a machine translated ouput, and when translating a previously created AD, and
(c) to research how end users would receive a text-to-speech voice in AD compared to a natural voice.

All experiments in the project departed from a single input, that is the film *Closer* (Nichols, 2004), because it had all the necessary materials available to carry out the quality evaluations.

3.1 Speech Recognition in Transcribing Audio Descriptions

This part of the project aimed to propose a process to automatically extract and transcribe the AD track from a movie using existing resources. The specificities of the process are described in Delgado *et al.* (forthcoming), and summarised below.

First, the movie soundtrack was extracted from the video file and converted to an adequate format, and the two available audio channels were mixed into a single mono channel. Then, downsampling was performed in order to obtain a 16 KHz, 16-bit, PCM wave file, generating a file containing both the movie soundtrack and the AD mixed together.

Secondly, an audio segmentation of the wave file was produced in order to keep exclusively speech content. This Speech Activity Detection (SAD) process was carried out with the acoustic segmentation tool included in the ALIZE toolkit (Fredouille *et al*., 2009).

Thirdly, the AD units were extracted from the audio track. A speaker model trained on the describer's voice could not be used because no training data were available, hence unsupervised approaches were followed: a speaker diarization based on the Binary Key

speaker modelling (Delgado *et al*., 2014) was performed over the speech signal output by the SAD module, the result being a text file that contained information about the detected speaker-homogeneous segments. For every segment, this included a speaker ID, a time-code in and a time-code out. Different speakers were detected and assigned a unique abstract identifier.

Fourthly, the abstract ID corresponding to the describer was identified manually. The obtained segments were processed to improve speech recognition results: segments less than one second long were discarded, close segments with a separation inferior to one second were merged, and an increase of 0.5 seconds both at the beginning and at the end was implemented to all segments.

Finally, these segments were used to split the signal into AD units, and the rest of speech was not taken into account. Each AD unit was isolated in an individual wave file. Next, the AD sound files obtained were automatically transcribed.

Although the speaker diarization process was carried out in two language versions of the movie (original English language, and dubbed version into Catalan), the transcription was only done in English using two automatic SR systems: (a) a large vocabulary continuous speech transcription system, tailored to achieve quality transcriptions of broadcast news audio, and trained on broadcast news audio and text (system A), and (b) a commercial dictation system trained for single speaker dictation purposes (system B).

Diarization Error Rates (DER) for speaker diarization were 22.6 in Catalan and 21.03 in English. Word Error Rates (WER) for the speech recognition tests were 64.43 for system A and 47.18 for system B. Missed speech time was the main error in DER (18.7 in Catalan, 11.8 in English), as there was high sound variability in the film, speakers talking under many acoustic conditions. Concerning SR, system performance was low due to the mismatch between the training conditions of the systems and the test materials.

All in all, these initial experiments have shown how speaker diarization is a necessary tool to isolate the describer voice as a previous step before SR implementation, while highlighting the potential and limitations of speech recognition. It remains to be seen what results would be obtained if engines were trained with specific corpora, a necessary step in future research.

3.2 Machine Translation in Audio Description

The second technology that was implemented in the process of AD was machine translation. The aim was to compare three situations: creation of AD, as it is standard practice, translation of an existing AD (from English into Catalan), and post-editing of a machine translated AD (from English into Catalan).

A necessary step was selecting the machine translation engine, hence a pre-test was carried out (Fernández-Torné and Matamala, 2014). Five professional translators volunteered to take part in the test. A clip from the movie *Closer* was selected, with an AD density of 240 words (1,320 characters distributed among 14 different AD units in 3.09 minutes). The excerpt was translated from English into Catalan using five free online machine translation engines, as the aim was to use existing free resources. The post-editing tool PET (Aziz *et al*., 2012) was customised for the experiment. Each participant was asked to post-edit five raw machine-translated versions of the excerpt in a randomized order. After post-editing each unit, participants were asked to evaluate various elements, indicating their level of agreement or disagreement with a given statement on a 5-point Likert scale. PE difficulty (De Sousa *et al*., 2011), PE necessity (Federmann, 2012), MT adequacy (Chatzitheodorou and Chatzistamatis, 2013), and MT fluency (Koehn and Monz, 2006; Koponen, 2010) were evaluated. Additionally, PE time and HTER were computed automatically (Specia, 2011). Finally, a ranking task was proposed to participants: they had to rank the translators from five (best) to one (worst) in a customised interface. A post-questionnaire provided more data on subjective

opinions, and HBLEU (Del Pozo, 2014) was also calculated automatically. All these indicators allowed us to choose the best machine translation engine freely available on the Internet for the purposes of our experiment (Fernández-Torné, forthcoming).

Once the engine had been selected, the main experiment took place. A homogeneous sample of 12 translators trained in AD were instructed to create an AD for three excerpts using three different approaches: (a) creating an AD *ex novo*, (b) translating and adapting, if necessary, an English AD into Catalan, and (c) post-editing the Catalan machine translation of an English AD generated by the engine selected in the pre-test. All excerpts were equivalent and tasks and clips were randomized across participants.

Participants were received in a computer lab, and then watched the entire movie. They were then asked to perform the three tasks using Subtitle Workshop, since this software allows to enter the time-codes. Input Log recorded all keyboard movement and time spent on each task. Pre-questionnaires and post-questionnaires gathered additional data, including subjective opinions on perceived effort. Keyboard logging allowed temporal effort, technical effort, and cognitive effort to be measured (Krings, 2001).

Results indicate no statistical differences among the three tasks in terms of temporal effort. Concerning technical effort, AD creation implies significantly more keyboard action than post-editing, and both AD creation and AD translation imply a higher number of characters typed than in the post-editing task. However, both AD translation and MT AD post-editing present a significantly greater number of mouse scrolls than AD creation. Cognitive effort is statistically higher in the AD creation task.

3.3 Text-to-speech in Audio Description

The aim of these experiments was to compare the reception of AD voiced by humans and voiced by text-to-speech technologies. A first test (Fernández-Torné and Matamala, 2015) was carried out to select the voices to be used in the main experiment. Twenty voices (5 male artificial, 5 male natural, 5 female artificial, 5 female natural) were used to record a random selection of AD units from the same stimuli, the film *Closer*. 20 participants assessed each voice using a five-point Likert scale on the following items, inspired by previous research (ITU, 1994; Viswanathan and Viswanathan, 2005; Hinterleitner *et al*., 2011, Cryer *et al*., 2010): overall impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, and acceptance. Two different lab sessions (one for artificial voices, one for natural voices) were done to avoid fatigue, and materials were randomized across participants. The results of these experiment allowed us to select the voices for the main test: two human voice talents, and two artificial voices (Laia by Acapela, and Oriol by Verbio).

The main experiment aimed to compare artificial and natural voice reception in AD by blind and low vision participants. 67 volunteers took part in the test. They listened to four randomized voices and responded to a questionnaire for each voice. Two different clips, equivalent in terms of length, intervening characters, background music, offensive content, and AD density, were used, one for female voices and one for male voices. This choice aimed to avoid participants' fatigue. Questionnaires assessed the same items as in the pre-test (see previous paragraph), plus additional subjective data. A statistical analysis was performed on quantitative data, showing that natural voices have statistically higher scores than artificial voices in all items under analysis. However, it is worth pointing out that no mean score of any of the items under analysis goes under 3.1 on a 5-point scale. For instance, the lowest value for the acceptance item is 3.7 (male text-to-speech) and the lowest score for overall impression is 3.2. (male text-to-speech). Additionally, 94% of participants state that text-to-speech AD is an "alternative acceptable solution" to human-voiced AD, and 20% of the

participants actually state that their preferred voice from the four included in the test is a synthetic one.

4 Conclusions

This project, exploratory in nature, has provided some innovative research in the field of audiovisual translation, where technological research has traditionally not been extensive until recently. It has focused on three technologies as applied in two genres and types of audiovisual transfer modes, providing new insights in how these technologies would affect not only the final product but mainly the key agents in the process (translators/describers) and also end users. However, some limitations must be acknowledged, due to the small scale of the project. A major setback is the low number of informants in many of the experiments, as well as the fact that the materials used in the experiments were not full programmes but just excerpts. For practical reasons, longer experimental sessions were not possible in a lab environment. Wider samples, ideally including professionals working with longer translations, are needed to shed more light on this topic which undoubtedly merits more research.

Acknowledgments

The research presented is part of the ALST project, funded by the Spanish Ministerio de Economía y Competitividad, reference code FFI2012-31024. Anna Matamala is also a member of the research group TransMedia Catalonia, funded by the Catalan Government (reference 201400027). The project has been possible thanks to eight researchers from UAB and six external participants, with special emphasis on the work carried out by Carla Ortiz-Boix and Anna Fernández-Torné as part of their PhDs. Thanks are also due to industries and end users cooperating in the various tests.

References

Aziz, Wilker, Sheila Castillo Maria de Sousa, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 3982-3987.

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.

Cryer, Heather, Sarah Home, and Sarah Wilkins, M. 2010. *Synthetic Speech Evaluation Protocol*. Technical report #7, Birmingham: RNIB Centre for Accessible Information (CAI).

Daniluk, Lukasz, Anna Matamala, and Pablo Romero-Fresco. 2015. Transcribing Documentaries: Can Respeaking Be Used Efficiently? Paper presented at the 5th International Symposium Respeaking, Live Subtitling and Accessibility, Rome.

De Sousa, Sheila Castillo Maria, Wilker Aziz, and Lucia Specia. 2011. Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 97–103.

Del Pozo, Arantza. 2014. *SUMAT Final Report.* http://www.sumat-project.eu/uploads/2014/07/D1-5_Final-Report-June-2014.pdf [last accessed September 14, 2015].

Del Pozo, Arantza, editor. 2013. *SUMAT: An Online Service for Subtitling by Machine Translation. Annual Public Report.* http://cordis.europa.eu/fp7/ict/language-technologies/docs/sumat-annual-report-2012.pdf [last accessed September 14, 2015].

Delgado, Héctor, Corinne Fredouille, and Javier Serrano. 2014. Towards a Complete Binary Key System for the Speaker Diarization Task. In *Interspeech 2014, Proceedings of the 15th Annual Conference of the International Speech Communication Association*, pages 572-576.

Delgado, Héctor, Anna Matamala, and Javier Serrano. Forthcoming. Speaker Diarization and Speech Recognition in the Semi-Automatization of Audio Description: An Exploratory Study on Future Possibilities. *Cadernos de Traduçao*.

Díaz-Cintas, Jorge, and Pilar Orero. 2006. Voice-Over. In Keith Brown, editor-in-chief, *Encyclopedia of Language & Linguistics*. Elsevier, Oxford, pages 477-479.

Federmann, Christian. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98: 25–35.

Fernández-Torné, Anna. Forthcoming. Machine Translation Evaluation through Post-Editing Measures in Audio Description.

Fernández-Torné, Anna, and Anna Matamala. 2014. Machine Translation and Audio Description. Is it Worth It? Assessing the Post-Editing Effort. Paper presented at Languages and the Media. 10th International Conference on Languages Transfer in Audiovisual Media, Berlin.

Fernández-Torné, Anna, and Anna Matamala, 2015. Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan. *The Journal of Specialised Translation*, 24: 61-88.

Franco*,* Eliana, Anna Matamala, and Pilar Orero. 2010. *Voice-over Translation: An Overview*. Peter Lang, Bern.

Fredouille, Corinne, Simon Bozonnet, and Nicholas Evans. 2009. The LIA- EURECOM RT'09 Speaker Diarization System. Paper presented at *RT'09, NIST Rich Transcription Workshop*. Florida, USA. http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/LIA-EURECOM_paper.pdf [last accessed September 14, 2015].

Hinterleitner, Florian, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock, C. 2011. An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks. In *Proceedings of the Blizzard Challenge Workshop*, *International Speech Communication Association*.

ITU-T Recommendation P.85 1994 *Telephone Transmission Quality Subjective Opinion Tests. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices.* ITU, Geneve.

Koehn, Philip, and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

Koponen, Maarit. 2010. Assessing Machine Translation Quality with Error Analysis. *MikaEL: Electronic Proceedings of the KäTu symposium on translation and interpreting studies*, 4. http://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf [last accessed September 14, 2015].

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

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014a. Cognitive Demand and Cognitive Effort in Post-Editing'. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice*, pages 73-84.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie. 2014b. Real Time Adaptive Machine Translation for Post-Editing with cdec and TransCenter. In *Proceedings of the Workshop on Humans and Computer-assisted Translation* (HaCaT), pages 72-77.

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

Lommel, Arle Richard, Alojscha Burchardt, and Hans Uszkoreit. 2013. *Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality*. ASLIB. http://www.mtarchive.info/10/Aslib-2013-Lommel.pdf [last accessed September 14, 2015].

Maszerowska, Anna, Anna Matamala, and Pilar Orero. 2015. Audio Description. New Perspectives Illustrated. Benjamins, Amsterdam.

Matamala, Anna, Pablo Romero-Fresco, and Lukasz Daniluk. Forthcoming. An Exploratory Study on the Application of Respeaking in the Transcription of Non-fictional Genres.

Orero, Pilar, and Anna Matamala. 2007. Accessible Opera: Overcoming Linguistic and Sensorial Barriers. *Perspectives. Studies in Translatology,* 15(4): 262-277.

Ortiz-Boix, Carla. Forthcoming. Post-Editing Wildlife Documentaries: Challenges and Possible Solutions. *Hermeneus*.

Ortiz-Boix, Carla, and Anna Matamala, Forthcoming a. Post-Editing Wildlife Documentary Films: a New Possible Scenario? *Perspectives. Studies in Translatology*.

Ortiz-Boix, Carla, and Anna Matamala. Forthcoming b. Quality Assessment of Post-edited versus Translated Wildlife Documentary Films: a Three-Level Approach. In *Proceedings of the Fourth Workshop on Post-editing Theory and Practice*.

Ortiz-Boix, Carla, and Anna Matamala. Forthcoming c. Assessing the Quality of Post-edited Wildlife Documentaries.

Rehm, George, and Hans Uszkoreit, editors. 2012. *Strategic Research Agenda for Multilingual Europe*. Springer, Berlin.

Romero-Fresco, Pablo. 2011. *Subtitling Through Speech Recognition: Respeaking*. St. Jerome, Manchester.

Sohn, Shara D. 2004. *Court Reporting: Can It Keep Up with Technology or will it be Replaced by Voice Recognition or Electronic Recording?* Honors Theses. Paper 265. opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1264&context=uhp_theses [last accessed 15 May 2015].

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*, pages 73–80.

Szarkowska, Agnieszka. 2011. Text-to-Speech Audio Description: Towards Wider Availability of AD. *The Journal of Specialised Translation*, 15: 142-162.

Viswanathan, Mahesh, and Madhubalan Viswanathan. 2005. Measuring Speech Quality for Text-to-speech Systems Development and Assessment of a Modified Mean Opinion Score (MOS) Scale. *Computer Speech and Language*, 19: 55-83.

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

Walczak, Agnieska, and Agnieska Szarkowska. 2012. Text-to-speech Audio Description of Educational Materials for Visually Impaired Children. In Silvia Bruti and Elena Di Giovanni, editors, *Audio Visual Translation across Europe: An Ever-Changing Landscape*. Peter Lang, Bern, pages 209-234.

Filmography

Nichols, M., director. 2004. *Closer*. Columbia Pictures, United States.

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