Universitat Autònoma de Barcelona

Audio description and tecnologies

Anna Fernández-Torné

Universitat Autònoma de Barcelona

Departament de Traducció i d'Interpretació

i d'Estudis de l'Àsia Oriental

Doctorat en Traducció i Estudis Interculturals

Audio description and technologies

Study on the semi-automatisation of the translation and voicing of audio descriptions

PhD thesis submitted by:

Anna Fernández-Torné

Supervised by:

Dr. Anna Matamala

2016

To Eva and David

Acknowledgements

There are many people whom I would like to thank warmly for their help and support.

To Iola Ledesma and Jose Navarro, from Escola Catalana de Doblatge (ECAD), for their cooperation in the recruitment of voice talents and the professional recording of all voice samples. And, of course, to all voice talents who voluntarily recorded the speech samples for my experiments.

To Anna Vilarnau, from iSpeech, for her readiness to provide me with their female Catalan voice and her interest in my research.

To Maite Caramazana, from the Àrea de Serveis de Difusió de la Subdirecció General d'Informació i Difusió Estadística in IDESCAT, and to Tomàs Jorge i Lacarta, from the Gabinet Tècnic de la Secretaria General del Departament de Benestar Social i Família from the Catalan Government, for providing me with such accurate data on my target population.

Special thanks to all the blind associations that, without expecting anything in return in the short term, recognised the future value of my work. In particular, to Manel Martí, Ainhoa, Felipe, Montse and Dani, from Associació Discapacitat Visual Catalunya (ADVC); to Pepa Casas, Meritxell Aymerich, José Ángel, Sabina and José Juan, from Associació Catalana per a la Integració del Cec (ACIC); to Adriana Pérez, from Organización Nacional de Ciegos Españoles (ONCE); to Víctor Corrales, from Associació Prodisminuïts de la UAB (ADUAB); and to Elías Barco, from Alternativa Social.

To all blind and partially sighted persons who participated in my tests, for their patience, their encouraging words and their engagement. My most sincere thanks. This research is for you.

To all other sighted participants who invested their valuable time and expertise in my small contribution to science, in particular to the 2013-2014 MTAV Catalan students, who proactively engaged in my MT experiment.

In the academic sphere, to Pilar Orero, for her innovative contributions and her cutting-edge ideas, and to Cristóbal Cabeza-Cáceres, Nazaret Fresno and Carla Ortiz Boix, with whom I have *mentally* shared part of this long journey.

To my beloved mum and dad, to my irreplaceable sister Maria, and to my dear friends, Maria and Yolanda, who have unconditionally supported me and believed in me no matter what. You are always there for me.

To you, Juan Carlos, who have shared with me so many ups and downs.

But most of all, I would like to express my immense gratitude to my supervisor, Dr. Anna Matamala, without whom I would surely not have embarked on this study. For your excellent academic leadership, your unvaluable guidance, your kind support and your warm understanding of all the vicissitudes that have occurred along these years. Thank you so much.

This project has been developed within the framework of the TransMedia Catalonia research group (2014SGR0027) and of the ALST project, funded by the Spanish Ministerio de Economía y Competitividad under grant number FFI2012-31024.

Table of contents

Annex 6.24. Masculine synthetic voice example sample

Annex 7 (electronic): TTS AD clips

Annex 7.1. Clip 1, feminine natural voice

Annex 7.2. Clip 2, feminine natural voice

Annex 7.3. Clip 1, feminine synthetic voice

Annex 7.4. Clip 2, feminine synthetic voice

Annex 7.5. Clip 1, masculine natural voice

Annex 7.6. Clip 2, masculine natural voice

Annex 7.7. Clip 1, masculine synthetic voice

Annex 7.8. Clip 2, masculine synthetic voice

Annex 8 (electronic): MT pre-test clip

Annex 8.1. Clip 1, Catalan dubbed version

Annex 8.2. English AD subtitles

Annex 9 (electronic): MT experiment clips

Annex 9.1. Clip A

Annex 9.2. Clip B

Annex 9.3. Clip C

Index of Tables

Index of Figures

Acronym Glossary

ACCEPT	Automated Community Content Editing PorTal
ACIC	Associació Catalana per a la Integració del Cec
AD	Audio description
ADUAB	Associació Prodisminuïts de la Universitat Autònoma de Barcelona
ADVC	Associació Discapacitat Visual Catalunya
AENOR	Asociación Española de Normalización y Certificación
ALST	Accessibilitat lingüística i sensorial: tecnologies per a les veus superposades i l'audiodescripció
APR	Average Pause Ratio
ARSAD	Advanced Research Seminar on Audio Description
AST	Audio subtitling
AV	Audiovisual
AVT	Audiovisual Translation
BLEU	Bilingual Evaluation Understudy
CAT	Computer-Assisted Tool
CCMA	Corporació Catalana de Mitjans Audiovisuals
CD-ROM	Compact Disk - Read Only Memory
CNGL	Centre for Next Generation Localisation
DVD	Digital Versatile Disk
EBMT	Example-based machine translation
ECAD	Escola Catalana de Doblatge
EMMA	European Multiple MOOC Aggregator
EU	European Union
FAS	Fundació Autònoma Solidaria
GQ	General Questionnaire
H1	Hypothesis 1
H2	Hypothesis 2
H3	Hypothesis 3
H4	Hypothesis 4

HBLEU	Human-targeted Bilingual Evaluation Understudy
HTER	Human-targeted Translation Edit Rate
IC	Índice de confianza
ICD	International Classification of Diseases
IDESCAT	Institut d'Estadística de Catalunya
ITU	International Telecommunication Union
ITU-T	International Telecommunication Union, Telecommunication Standardization Sector
MA	Master
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MOS	Mean Opinion Score
MT	Machine translation
MUSA	Multilingual Subtitling of multimedia content
NC	North Carolina
NIST	National Institute of Standards and Technology
ONCE	Organización Nacional de Ciegos Españoles
OR	Odds ratio
PE	Post-editing
PET	Post-editing Tool
PIUNE	Pograma per a la Integració dels i les Universitàries amb Necessitats Especials
PQ	Profile Questionnaire
PTQ	Post-task Questionnaire
PWR	Pause to Word Ratio
RAM	Random-Access Memory
SMT	Statistical machine translation
Stdev	Standard Deviation
SUMAT	Subtitling by Machine Translation
TAUS	Translation Automation User Society
TECNACC	Tecnologia per a l'accessibilitat

TER	Translation Edit Rate
TM	Translation Memory
TrAID	Translation Aid
TTS	Text-to-speech
TV	Television
UAB	Universitat Autònoma de Barcelona
UK	United Kingdom
UN	United Nations
UNE	Una Norma Española
USA	United States of America
USB	Universal Serial Bus
VLC	VideoLan Client

Chapter 1. Introduction

1 Introduction

> 1. States Parties recognize the right of persons with disabilities to take part on an equal basis with others in cultural life, and shall take all appropriate measures to ensure that persons with disabilities:
> (a) Enjoy access to cultural materials in accessible formats;
> (b) Enjoy access to television programmes, films, theatre, and other cultural activities, in accessible formats;
> (c) Enjoy access to places for cultural performances or services, such as theatres, museums, cinemas, libraries and tourism services, and, as far as possible, enjoy access to monuments and sites of national cultural importance.
>
> 30th article of the UN's International Convention on the Rights of Persons with Disabilities (2006)

> The right of persons with a disability and of the elderly to participate and be integrated in the social and cultural life of the Union is inextricably linked to the provision of accessible audiovisual media services. The means to achieve accessibility should include, but need not be limited to, sign language, subtitling, audio-description and easily understandable menu navigation.
>
> Consideration number 46 of the Audiovisual Media Services Directive 2010/13/EU (2010)

Society's commitment towards accessibility is growing rapidly around the globe. Lately audiovisual contents are perhaps one of the areas in which most efforts are being invested due to their relatively recent growth and massive expansion. The contents of theatres, cinemas, museums, TV channels and web sites, among others, need to be made accessible to people with any kind of sensorial disability.

As far as blind and visually impaired persons are concerned, audio description (AD) is the technique used to turn visual information (images) into oral information. Therefore, it is used to make audiovisual contents accessible to blind and partially sighted people. However, AD can also be used with other purposes. The Spanish

Standardisation and Certification Association AENOR states in relation to the AD standard UNE 153020 that "the criteria […] apply to the AD targeted at blind and visually impaired persons", but it adds that "people with no visual disabilities can also benefit from it" (AENOR, 2005, p. 4). It can therefore be deduced that there will be two different sets of users. On the one hand, there are those who depend to a great extent on AD to fully comprehend the film and for whom AD is an essential resource to entirely enjoy it. This would be the case of blind and visually impaired people, the main AD target audience. On the other hand, there are those who just use AD as an extra support to thoroughly appreciate it, such as elderly people, with decreasing sensorial or cognitive abilities; people who cannot or do not want to see the images for any particular reason in a given moment; or even people who are learning a second language and to whom AD provides new vocabulary directly related to images.

AD users have been increasing in the last few years due to population aging. If we take into account that most visually impaired people fall into the age group comprising from age 65 onwards (see Table 1.1), we should expect a sustained increase in the number of AD users for the subsequent decades, as 23.55% of the EU population is expected to be 65 years old or more in 2030 (see Table 1.2). Taking Catalonia as an example, the number of people with legal visually impairment has raised considerably since 2002 (see Figure 1.1).

	From 6 to 64	**From 65 to 79**	**From 80 onwards**	**Total**
Spain	306.000	337.200	336.000	979.200
Catalonia	34.800	44.300	44.600	123.700

Table 1.1. Population with visual impairment by age group in Spain and Catalonia in 2008, according to the Spanish National Statistics Institute

	Population	%
From 0 to 15	80879	15.56
From 16 to 64	316598	60.89
From 65 onwards	122465	23.55
Total	519942	

Table 1.2. Projected population in millions of inhabitants on 1st January by age groups in the EU for the year 2030, according to Eurostat

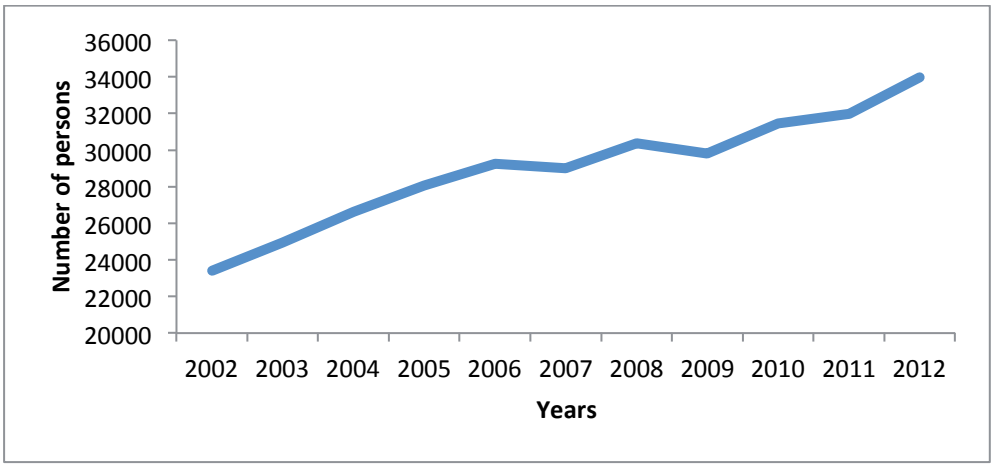


Figure 1.1. Persons legally recognised as being visually impaired in Catalonia, according to Idescat

Consequently, there is an urge for an immediate increase in the number of AD products and services, if the rights of a raising number of people are to be ensured. However, this technique entails a laborious process which hampers desirable levels of audiovisual products to be audio described due mainly to two factors: the time that the process of AD script creation and voicing entails, and the costs associated to their production.

In view of this need, this thesis explores the application of technologies to audio description in order to optimise the present AD workflow. On the one hand, the focus will be on the voicing of ADs by means of text-to-speech (TTS) systems as an

alternative to human-based systems of delivery. On the other, the interest will lie in the inclusion of machine translation (MT) in the process of creation.

1.1 Objectives and hypotheses

This research has two main objectives, which in turn have other specific goals associated:

1. to study the reception of Catalan text-to-speech audio descriptions by blind and visually impaired persons, compared to traditional human-voiced audio descriptions.
2. to research the inclusion of English-Catalan machine translation systems in the process of AD creation, compared to the process of human creation and human translation, focussing specifically on the effort involved in each process.

In order to attain these objectives, some additional specific aims need to be fulfilled.

Regarding the first general objective, linked to text-to-speech audio description, the specific aim is the following:

1. to assess a selection of available synthetic and natural voices in Catalan, in order to determine which are the most adequate ones for the purposes of our research;

In order to fulfil this aim, a methodological framework for the evaluation of voices will need to be developed, as explained in chapter 2.

Regarding the second general objective, linked to machine translation audio description, the specific aim is:

1. to assess a selection of available free online machine translation engines from English into Catalan in order to determine which is the most adequate for our purposes in the audio description domain;

In order to select the machine translation engines, a methodological framework will also need to be established, as described in chapter 2. Additionally, our theoretical concept of “effort” as well as the methodological tools measuring it will need to be put forward (see chapter 5).

Our working hypotheses in relation to the objectives specified above are expounded next.

H1. Blind and visually impaired persons will accept text-to-speech audio descriptions.
H2. Blind and visually impaired users will prefer human voiced audio descriptions compared to text-to-speech audio descriptions.
H3. The effort of post-editing a machine translated audio description will be lower than that of creating an audio description from scratch.
H4. The effort of post-editing a machine translated audio description will be lower than that of translating it manually.

1.2 Thesis structure

This is a thesis by publications, also known as a compilation thesis. As such, it has been written in the format of a compendium of interrelated articles. It is composed of a chapter per article plus four additional chapters: an introductory chapter explaining the connection between the articles and presenting the thesis as a unified whole; a methodological chapter, where the methodological approach is put forward and the articles are presented; a summary chapter, in which the thesis is outlined; and a concluding chapter, which consists of a discussion of the results and the final conclusions.

The final structure of the thesis is described in more detail next.

Chapter 1 sets forth the objectives and hypotheses of our research (section 1.1), explains the structure of the thesis (section 1.2) and exposes the theoretical framework and previous relevant research related to the topic developed in the thesis

(section 1.3). Even though this last section is not formally required in a thesis presented as a compendium of publications, we have deemed it necessary since the regular article format does not allow for a comprehensive presentation of the general framework in which our research is to be embedded.

Chapter 2 presents an introduction of the methodology applied to meet the thesis' objectives. Although each article refers to an experiment and describes its specific methodology, we have included a summary of the methodology in a unified way. Again, even though this chapter is not formally required in a thesis by publications, it has been included for the sake of clarity and coherence. Thus, this chapter describes the methodology used in two studies, each of them subdivided into a pre-test and a main experiment:

a) Methodological aspects on the application of semi-automated processes to the voicing of audio descriptions through text-to-speech (section 2.1).
    i. Pre-test for the selection of the voices to be used in the main experiment (subsection 2.1.1).
    ii. Experiment on the reception of text-to-speech audio descriptions compared to human voiced ones (subsection 2.1.2). This experiment corresponds to article 1, included in chapter 3.
b) Methodological aspects on the application of semi-automated translation processes to the creation of audio descriptions through machine translation and post-editing (section 2.2).
    i. Pre-test for the selection of the machine translation engine to be used in the main experiment (subsection 2.2.1). This corresponds to article 2, included in chapter 4.
    ii. Experiment on the comparison of audio description creation, translation and post-editing efforts (subsection 2.2.2). This corresponds to article 3, included in chapter 5.

Chapter 3 includes article 1: Fernández-Torné, A., Matamala, A. (2015). Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan. *The Journal of Specialised Translation*, 24, 61-88. This article presents a user study conducted with 67 blind and partially sighted people who assessed two synthetic and two natural voices in AD. The aim was to determine whether blind and visually impaired people would accept the implementation of TTS in the AD of dubbed feature films in the Catalan context.

Chapter 4 corresponds to article 2: Fernández-Torné, A. (Forthcoming). Machine Translation Evaluation through Post-Editing Measures in Audio Description. *inTRAlinea*. 2016, 18. 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. The best-rated engine would therefore yield the most suitable freely machine-translated audio descriptions in Catalan, presumably reducing the audio description process turnaround and costs.

Chapter 5 presents article 3: Fernández-Torné, A., Matamala, A. (2016). Machine Translation in Audio Description? Comparing Creation, Translation and Post-editing Efforts. *Skase*, 9(1), 64-85. This article presents an experiment comparing the efforts of creating an audio description from scratch, translating it manually from English into Catalan and postediting its machine translated version. The results of this experiment have shown that the objective post-editing effort is lower than creating it from scratch.

Chapter 6 includes a brief summary of the thesis, condensing the most relevant information, as required to all theses by publications.

Chapter 7 is the concluding chapter and comprises a discussion of the results together with the final conclusions of the study on text-to-speech (section 7.1) and of the study on machine translation (section 7.2), and the possible paths for future research (section 0).

Next, the Updated bibliography is provided, unifying the different citation styles under one common format and offering the up-to-date versions of the references cited in the articles some time ago. Finally, the Annex are presented. These include:

- Annex 1: Articles as published
- Annex 2: Documents related to the TTS pre-test for the selection of best-rated voices
- Annex 3: Documents related to the TTS main experiment on the reception of TTS AD
- Annex 4: Documents related to the MT pre-test for the selection of a MT engine
- Annex 5: Documents related to the MT main experiment on the comparison of AD creation, translation and PE efforts
- Annex 6 (electronic): Voice samples
- Annex 7 (electronic): TTS AD clips
- Annex 8 (electronic): MT pre-test clip
- Annex 9 (electronic): MT experiment clips

1.3 Theoretical framework and research background

> As an interdiciplinary area of research, translation studies attracts students and scholars with a wide range of backgrounds, who then need to face the challenge of accounting for a complex object of enquiry that does not adapt itself well to traditional methods in other fields of investigation.
>
> Saldanha and O'Brien (2013, p. v).

According to Remael, Reviers, and Vercauteren (2015) AD is a service that "offers a verbal description of the relevant (visual) components of a work of art or media product, so that blind and visually impaired patrons can fully grasp its form and content" (ibid., p. 9). Arma (2011) defines audio description as a "technique aimed at enhancing the accessibility of different types of audiovisual products primarily to the blind and the visually impaired, using a pre-recorded or live audio track which is

inserted into non-meaningful pauses within dialogues in order to 'translate' into words the visual elements otherwise only accessible to sighted users" (ibid., p. 10). This "translational" dimension sets audio description within Translation Studies, the broad theoretical framework in which this PhD thesis is to be embedded, focusing on the purpose of the translation while adapting "the text according to the needs of the future readers" (Suojanen, Koskinen, & Tuominen 2015, p. 1). In particular, audio description is considered an audiovisual translation modality (Maszerowska, Matamala, & Orero, 2014), which helps narrow down the scope of the broader theoretical framework into Audiovisual Translation and Media Accessibility Studies (Remael, Orero, & Carroll, 2012).

Translation process research and reception research in audiovisual translation studies have also nurtured the theoretical background of our thesis. Saldanha and O'Brien (2013) classify research according to where the focus of attention is, namely the translation product, the translation process, the participants involved in the process and its context. In relation to participants, they distinguish between participants as agents directly taking part in the process and participants as respondents answering a questionnaire presented by the researcher.

According to the authors' classification of research, our empirical studies relating to MT gather observations on audio description as a translation process, as we collect objective and subjective user activity data (Carl, Jakobsen, & Jensen, 2008) through keylogging techniques and questionnaires. The MT studies also gather observations on translators, as active participants in the process. In contrast, the studies referring to TTS centre the attention on participants as end-users of the audio description.

The application of technologies to audio description is also of utmost importance in our research. Thus, other concepts related to the technologies applied, namely text-to-speech and machine translation, are most relevant and need to be included in our theoretical framework, such as synthetic speech evaluation, machine translation quality assessment and post-editing effort measurement. Since a definition and state

of the art of these concepts is thoroughly provided for in the articles, it will not be included here to avoid repetition. It will only suffice to say that effort is understood in this thesis as the amount of physical and/or mental exertion expended for a specified purpose, and is assessed following Krings' (2001) proposal. Krings differentiates between temporal, technical and cognitive effort. Temporal effort is the total time spent on post-editing a text, technical effort refers to the operations carried out to post-edit the text, and cognitive effort applies to the mental processes involved in identifying errors in raw machine-translated texts and in deciding on the necessary steps to correct them.

Thus, this section intends to give a general overview of the literature in which these technologies have been used in the audiovisual translation field. Additionally, since this research has adopted an empirical approach with a strong focus on reception, a brief account of the studies dealing with filmic audio description from a reception point of view will be provided.

It must be noted that it is not our purpose to offer an exhaustive review of studies devoted to other aspects of AD as its origins and evolution, or the existing standards and norms. Nor are we committed to analysing AD from a didactic, narratological, linguistic or filmic perspective, as other articles (Braun, 2008) and recent PhD thesis provide such information in a comprehensive way (Arma, 2011, Ramos, 2013, Cabeza-Cáceres, 2013, Fresno, 2014).

1.3.1 Audiovisual translation and text-to-speech

Blind and visually impaired persons were the one of the first target users for which the first text-to-speech systems were designed. In 1976 Raymond Kurzweil introduced the Kurzweil Reading Machines for the Blind, a reading aid system with an optical scanner to read aloud written text for the visually impaired (Lemmetty, 1999). Nowadays ther are many applications of speech synthesizers for the blind: from audiobooks and screen reading on computers and mobiles (Asakawa & Takagi, 2007), to mobility aids

(GPS navigation devices) and entertainment (audio subtitles and audio description) (Cryer & Home, 2008). All these applications are no longer just targeted at the blind and visually impaired, but at persons with other handicaps (e. g. with reading difficulties or cognitive impairments) and at sighted audiences too.

In the context of entertainment, audio subtitling (AST), "the spoken rendering of the written (projected) subtitles or surtitles with a filmed or live performance" (Remael, 2012), has introduced synthetic voices to automatically read aloud the subtitles and make them accessible. This service has been implemented in television broadcasts in countries such as the Netherlands (Verboom, Crombie, Dijk, & Theunisz, 2002), Sweden (De Jong, 2006), and the UK (Hailes, 2013). Also in 2013 the Catalan public television (CCMA, 2013) started offering TTS AST in their afternoon and evening news programmes to voice the Catalan overprinted subtitles of the original declarations of the characters intervening in the informations provided.

Nielsen and Bothe (2007) proposed a user-based device for reading aloud subtitles (SubPal). The device was applicable for television and in the cinema, while the fact that it used a multilingual speech synthesizer made it suitable for several countries. To expand the availability of spoken subtitles and avoid the need for a special decoder, Derbring, Ljunglöf, and Olsson (2009) developed a free and open-source tool within the SubTTS project to read Swedish subtitles aloud.

In Poland, Mączyńska (2011) tested the synthesised audio subtitling together with the TTS AD in a non-fiction film (*La Soufriere*, by Werner Herzog). The study involved 54 blind and partially sighted and 30 sighted participants. The artificial voices tested were Krzysztof (male voice) for the AST and Zosia (female voice) for the AD, both by Loquendo. This study was part of a wider project developed at the University of Warsaw, which analysed the application of text-to-speech audio description assessing its reception among blind and visually impaired persons of different ages in several types of audiovisual products, namely:

- in the dubbed educational animation series for children *Once Upon a Time… Life* (Walczak & Szarkowska, 2010), involving 76 blind and partially sighted participants;
- in the monolingual feature film in Polish *Dzień Świra* (*The Day of the Wacko*), by Marek Koterski (Szarkowska, 2011), involving 24 blind and partially sighted participants;
- in the dubbed feature film *Harry Potter and the Philosopher's Stone*, by Chris Columbus, (Drożdż-Kubik, 2011), involving 17 blind and partially sighted participants.
- in the foreign fiction film *Volver*, by Pedro Almodóvar, with voice-over (Szarkowska & Jankowska, 2012), involving 18 blind and partially sighted and 2 sighted participants;

The results of all 5 studies indicated that a majority of viewers accepted TTS AD.

Pazos (2012) also assessed the reception of the Spanish TTS AD of a humour sketch entitled *Para Elizabeth* by Les Luthiers. 20 blind and visually impaired persons participated in the study and the conclusions matched those of the Polish project, with a majority of participants accepting the AD synthesised using the voice Antonio, by Assistive Ware.

Likewise, Ortiz-Boix (2012) explored the application of the Spanish TTS to AD and tested the reception of four masculine synthetic voices, two feminine synthetic voices and one natural voice with no distortion, a few distortion and a lot of distortion with 18 sighted participants. Results showed that some synthetic voices would not be accepted to voice ADs (such as the masculine synthetic voice by Festival), while others definitely would (masculine synthetic voices by Acapela and Verbio) and were preferred to the distorted natural voices.

In the context of Internet videos' accessibility, Chapdelaine and Gagnon (2009) presented a website platform that offered end-users with two levels of AD rendered by a synthetic voice, although the language is not stated in the article. The evaluation

questionnaire was aimed at providing feedback on the participants' interaction with the player and the website, and with the AD's contents and delivery. Results indicated that the lowest scores dealt with the quality of the synthetic voice.

Similarly, Kobayashi, Fukuda, Takagi, and Asakawa (2009, p. 249) developed a "technique to use synthesized speech to add ADs to online videos on any websites". The first step of their project included determining whether or not synthesized voices could compete with real voices. In relation to TTS, three kinds of voice were tested (human, standard TTS, and prototype TTS), including additional variables such as expressive TTS technology vs standard TTS at a later stage. Again, this broad reception study performed in Japan and in the USA concluded that both Japanese and English TTS AD was generally accepted, especially for relatively short videos and informational content.

In 2011, the company Apps4Android designed and developed "a high-quality, Android-based, text-to-speech driven, application that provides the ability to add audio-descriptions to You Tube videos" (Jacobs, 2011). Based mainly on educational video tutorials, the application used English TTS technology to voice the audio descriptions pausing the video instead of limiting narrations to the natural pauses in the audio.

Also in that year, Mieskes and Martínez (2011) presented a workplace for transforming text into speech. Authors proposed the voicing of audio descriptions as one of its main applications. The editor allowed choosing the voice, the speaking rate and the pitch, and had phonetic tuning functionalities. The user may upload an existing description, translate it or create a new one, upload the corresponding movie and synthesise the description. Although first focused on German, the authors aimed at developing a multilingual system.

Encelle, Ollagnier-Beldame, Pouchot and Prié (2011) presented an exploratory work on video accessibility on the web for the blind and visually impaired with audio enrichments "focused on the use of earcons as a way to complement speech synthesis for conveying visual information" (ibid., p. 129). Even though they do not specify the

language of the TTS system tested, their study showed that earcons combined with speech synthesis enhanced the understanding of set-related information, although several participants asked for greater conciseness.

1.3.2 Audiovisual translation and machine translation

As far as the application of MT to audiovisual content is concerned, research has been basically focused on the subtitling field. Popowich, McFetridge, Turcato, and Toole (2000) studied for the first time the automatic translation of closed captions from English into Spanish using a rule-based MT system and concluded that MT could already be satisfactorily applied to the subtitling domain.

O'Hagan (2003) sought to know if the available language technology could be applied to subtitle translation, for which she tested both "the applicability of TM [translation memory] in a context of serialized films where the subtitler could leverage the subtitles produced for the previous episode" and "the usability of freely available MT for creating subtitles mainly by non professional subtitlers" (ibid., p. 14). Results showed that the TM tool was of very little help in the translation, while most of the raw MT outputs of the English subtitles could be helpful for non-English speaking viewers to roughly understand what was going on in the film, implying that there was a clear scope for potential.

Piperidis, Demiros, and Prokopidis (2004) integrated machine translation (Systran) and translation memory (TrAID) in the translation component of the MUSA project, aimed at the generation of multilingual subtitles combining speech recognition, advanced text analysis and MT technologies for the English-French and English-Greek linguistic combinations. The acceptability of the translated subtitles resulted in the range of 45% to 55%, while the evaluation of the integrated prototype showed that it would entail considerable productivity gains (Piperidis et al., 2004).

In the context of the eTITLE project, Melero, Oliver, and Badia (2006) presented a web service that combined MT with TM technologies for the multilingual subtitling of

audiovisual content in several linguistic pairs, including English-Catalan, Spanish-Catalan and English-Spanish, together with the linguistic combination English-Czech. However, only the translation memory for the English-Czech combination was filled with subtitles, while the Spanish-Catalan and English-Spanish combinations used newspaper articles and texts from the Catalan Government's official bulletin, and UN texts respectively. Results showed that integrating TM and MT solutions obtained better results than the use of MT alone.

Armstrong, Caffrey, and Flanagan (2006) focused their research on an example-based machine translation (EBMT) system for the English-German and English-Japanese language combinations. They fed the system with existing human translations from DVDs to automatically produce subtitles and evaluated the output using automatic metrics (BLEU) and testing their reception with German and Japanese native speakers. They also evaluated the quality of automatically translated subtitles using EBMT systems trained on homogeneous and heterogeneous corpora for the same language combinations and results showed that the best rated subtitles were the post-edited EBMT ones before the versions by Babelfish and the unedited EBMT versions (Caffrey, 2006). They also tested the feasibility of using an EBMT engine trained on both homogeneous and heterogeneous data to produce subtitles for the language directions German-English and English-German. Results showed that the translation quality obtained was higher when training the engine on homogeneous data (Armstrong et al., 2006).

Volk (2008) explored the application of a 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 proved to be considerably shortened by the use of such a trained MT system.

De Sousa, Aziz, and Specia (2011) used a rule-based MT engine (Systran SMTU), an in domain SMT system trained with more than 74,000 sentence pairs aligned from the

fan subtitles from the TV series *X files*, an out-of-domain SMT (Google Translate) and a TM system (Trados Studio) to compare the effort involved in translating subtitles manually from English into Portuguese from scratch with the effort of post-editing (semi)automatically translated subtitles. Time was used as an objective measure for PE effort and results showed that post-editing was much faster than translating subtitles from scratch.

In the final report of the SUMAT project, which offered an online service for subtitling using statistic MT technologies covering nine European languages combined into seven language pairs in both directions, Del Pozo (2014) stated that the systems were trained with a combination of parallel professional and crowd-sourced subtitles. For the final versions of the statistic MT systems to be developed, two large-scale evaluations were performed based on quality rating and productivity gain/loss (Etchegoyhen et al., 2014). Quite positive results were obtained by both objective metrics and the ratings by professional users "with significant portions of MT output deemed of sufficient quality to reach professional quality standards through minimal to medium post-editing effort" (Del Pozo, 2014, p. 40). A global productivity gain of almost 40% was also reported.

In turn, the EU-BRIDGE project (2012-2015) focused on the creation of applications that made use of speech translation: from captioning translation for TV broadcasts, to automatic transcription and translation of webinars.

The EU has funded several projects dealing with the automatic generation of subtitles and their translation into multiple languages both in media –MUSA (2002-2004), eTITLE (2004-2006), SUMAT (2011-2014) and EU-BRIDGE (2012-2014)– and educational content –transLectures (2011-2014) and EMMA (2014-2016). However, the implementation of MT in the AD domain has not yet attracted the attention of many researchers, or the funding of any field-related project.

To the best of our 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, and Apertium was used as an example of rule-based engine. 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 (Matamala, 2015). This thesis is also part of the ALST project, which researches the application of three technologies, including speech recognition, machine translation and speech synthesis, to two oral modes of audiovisual translation, namely voice-over and AD.

1.3.3 Studies on the reception of audio description

Although reception studies in the audio description field research date back to the 90's, they have proliferated in the last decade. The emphasis has been usually placed on the main target users of AD, namely blind and visually impaired persons, whose perception and opinions are central for the development and improvement of all aspects involved in the audio description process and final product. However, some studies have focused on the reception of audiovisual products by sighted persons with the aim to analyse how they perceive and process the information contained in audiovisual texts (Mazur & Chmiel, 2012; Orero & Vilaró, 2012; Kruger, 2012). In spite of this, this thesis will only cater for those studies centered on the blind and visually impaired, which constitute our ultimate target audience.

Moreover, as our research revolves around the audio description of feature films, studies on the reception of the audio description of the scenic arts (Matamala, 2005; Orero, 2007a; Weaver, 2013; Udo & Fels, 2009) will also be omitted. For the reception studies relating to TTS AD, please refer to subsection 1.3.1.

In 1991 the Royal National Institute for the Blind (RNIB) conducted the AUDETEL (Audio Described TELevision) project in the UK. The project's aim was to enhance the television for visually impaired people. Thus, it collected "the TV viewing habits, difficulties and preferences of about 200 visually impaired informants" (Braun, 2008).

Peli, Fine, and Labianca (1996) explored the reception of short segments of two TV programmes without AD. They presented them to 25 visually impaired participants and to 24 sighted participants, while 29 other sighted participants only heard the audio of such TV programme segments. They all had to answer questions related to visual information contained in the AD and results indicated that without hearing the AD "the subjects with normal vision performed the best, followed by those with low vision and those who heard only the audio portion" (Peli et al., 1996, p. 378), which suggested that AD actually helped the blind and visually impaired to access visual details that they would otherwise miss.

Schmeidler and Kirchner (2001) conducted an experiment with 111 legally blind participants and results showed that the participants watching TV science programmes with AD obtained more information and had more enjoyment than those participants watching them with no AD.

As reported by Ramos (2013), Herrador Molina studied the reception of British AD scripts translated into Spanish by blind and partially sighted audiences and concluded that, contrary to other scholars' opinion (Rodríguez Posadas & Sánchez Agudo, 2008), they accepted and enjoyed the British AD style translated into Spanish.

Luque (2009) explored the reception of the metaphors in the Spanish AD scripts by blind and visually impaired audiences. She presented two film excerpts with two different ADs: one containing few or no metaphors, and another manipulated one incorporating metaphors. Ten blind and visually impaired participants answered two comprehension questionnaires corresponding to each of the excerpts and a common questionnaire on the style of the AD, their preferred AD version, their opinion on which gave a better grasp of the scene and anything they had particularly liked or

disliked in the AD. Results showed that the understanding of metaphors by end users did not imply any additional difficulty, and that the presence of metaphors favoured the understanding of the film images and the viewers' immersion in the film.

On behalf of the RNIB, Rai (2009) carried out a study by which the demand for audio described Bollywood films in the UK and India was explored. He used both quantitative and qualitative research. The first study involved 260 blind or partially sighted participants of Asian descent living in the UK, while the qualitative study involved fifty blind or partially sighted participants. The study showed that there was a "huge unmet need for audio described [Bollywood] films in the UK and India" (ibid., p. 3), with participants reporting an improvement of the film understanding thanks to AD. The study also identified that respondents lacked awareness about audio description, preferred that there was AD over the songs and favoured AD in Hindi rather than in English.

In their article, Chmiel and Mazur (2012) focused on the methodological issues that arouse when conducting a pilot reception study involving 18 participants, and the changes the results implied in the questionnaires developed for the research project AD-Verba, which aimed at collecting user information and preferences for the creation of AD guidelines in Poland. Their study consisted of face-to-face interviews, which comprised three questionnaires. The first one gathered information on the participants' demographics and usage of audio described products. The second one included 11 comprehension and/or preference questions related to the three AD clips they had presented to the participants and referring mainly to a moment in which the names of the characters should be introduced and the degree of explicitness of descriptions. Finally, the third questionnaire incorporated questions in relation to general AD preferences (on the gender of the voice talent reading the AD, on the AD objectivity degree, on the description of gestures), and opinions on the usefulness of AD for the comprehension of the film, on the AD delivery speed, pauses and overlaps with dialogues, and on the inclusion of elements such as evaluative adjectives, colours,

similes, logo, opening titles, credits, the name of the audio describer and the name of the voice talent.

Fryer and Freeman (2012) compared the reception of different styles of AD by 18 blind, 18 visually impaired and 25 sighted participants. They were presented with a clip of the film *Brief Encounter*, by David Lean, audio described in a standard way and audio described in a "cinematic" way, i.e. adding filmic language. Results showed a unanimous preference for the presence of AD, with 66.7% preferring the cinematic AD over the standard AD.

The same authors (Fryer & Freeman, 2013a) conducted another reception study involving 18 blind, 18 partially sighted and 18 sighted participants in which the levels of presence, understood as "the psychological sense of immersion in any mediated environment" (ibid., p. 15) for the same film clip with no AD, with "standard" AD and with "cinematic" AD were compared. Results showed that blind and partially sighted people were most engaged when watching the stimuli with the cinematic AD, also reporting higher presence levels with AD than sighted people with no AD.

In a study with 20 blind and visually impaired persons, Romero-Fresco and Fryer (2013) assessed the acceptance of cinematic audio introductions for a romantic drama movie and for a documentary. The results showed that most participants accepted such cinematic introductions and considered that the style was correct and that the amount of information was adequate.

In the context of the comprehension of audio described films, Cabeza-Cáceres (2013) carried out a reception study in which several parameters of the AD (speed, explicitation of information and intonation as independent variables) were modified to see if they had an impact on the comprehension (dependent variable) and enjoyment. There was a group of 30 blind and visually impaired participants and a control group of 10 sighted participants. The results showed that a. there was a negative correlation between the speed of the AD and comprehension; b. the intonation did not affect comprehension; and c. there was a positive correlation between AD explicitation and

comprehension. As for the enjoyment, speed and explicitation did not affect the enjoyment, whereas intonation (too uniform or too emphatic) had a negative influence on it.

Ramos (2013) researched on the emotional impact of 15 film excerpts entailing three emotions, namely disgust, fear and sadness. By means of two different studies involving 30 blind and visually impaired and 40 sighted participants, she first tested the emotional impact provoked by different modalities (original version films, films without images and audio described films) in order to elucidate whether words may cause an emotional impact similar to that of images. She also tested the emotional response provoked by two different AD in terms of the degree of objectivity. The results of the first study showed that the different emotions were transmitted through different degrees of visual component and thus, the higher the visual component, the higher the need of the AD. Hence, the disgust emotional response for the participants watching the original film was similar to that of the participants watching the audio described version, which did not happen in the case of fear and sadness due to the importance of music and dialogues in the conveyance of the emotions in such scenes.

Fresno (2014) explored the reception of audio described characters by blind and partially sighted persons. In particular, she tested how the amount of information included in the AD of characters and its presentation had an influence on blind and partially sighted users' recall. Involving 44 blind and partially sighted participants, the results showed that there was a negative correlation between the amount of information included in the AD and the amount of information correctly remembered. They also indicated that dividing the information into short units and delivering them at different stages helped improve users' memory.

Szarkowska and Jankowska (2015) reported on a qualitative exploratory study performed in Poland in the years 2010-2012 involving open screenings of several audio described foreign films after which discussions and interviews were held and questionnaires were given out to the blind and partially sighted audience. The authors

presented the main challenges of audio describing foreign films in a non-dubbing country where voiceover is used, and proposed several solutions to these challenges. The solutions included inserting names for the identification of speakers, and providing audio introductions and applying strategies already used in the translation field to the audio description, such as naming, explicitation, generalisation, specification and retention, to deal with the description of cultural elements and the problems posed by intertextuality. The results indicated that "the optimum synchronisation is when the original dialogue can be heard in the background and when AD does not overlap with the original actors' voices or with the voiceover translation" (ibid., p. 245) and also demonstrated the feasibility of applying translation strategies to audio descrtiption.

Finally, in the 2015 edition of the Advanced Research Seminar on Audio Description (ARSAD), Bardini presented her reception study which she was conducting in the broader context of her PhD thesis on the transposition of film language in AD. Although no results were given yet, the study was centered on the film experience of visually impaired persons. She tested three AD styles (a conventional version and two other verions including film language) to evaluate the quality of end users' film experience in each case and collect information on their preferences. Jekat, in turn, used the semantic differential to analyse quantitatively the meanings and connotations of affective words and compared the reception of audio described films by blind and partially sighted persons with the reception of the same films with no audio description by sighted participants. The results showed that perceptions by both audiences were quite similar when the descriptive attributes were semantically clear, while perceptions differed when more subjective or abstract attributes were used.

All in all, this thesis is fostered by this relevant research in AVT. First, in relation to TTS, it particularly follows in the wake of Kobayashi et al.'s (2009) work, in that they also compared the reception of human-voiced and TTS audio descriptions. Second, although mainly focused on the subtitling domain, the results of the MT works

presented justify one of the main objectives of this thesis, as it is more than reasonable to believe that the same advantages may be obtained when applying MT to audio description. Third, the literature in AD reception proves that it is a valuable source of information for researchers to assess whether they are on the right track as they receive first-hand feedback from end-users. However, many technical aspects in relation to the quality of the synthetic voices, cultural considerations and even the degree of use of technologies by certain groups of people come into play in this kind of tests (Cryer & Home 2009). Therefore, the Catalan context was found to be a gap in present research and thus the focus of our academic attention.

Chapter 2. Methodology

2 Methodology

> The methods traditionally used by literary scholars and Humanists are often merely speculative, in the sense that there are very few checks on the assertions made. Usually they are also of a rather subjective nature and tend to rely on methods such as introspection (subjectively tracking one's own thoughts). These approaches may have their benefits in some respects, but they do not yield particularly reliable forms of information. They generate more concepts, more theoretical constructions, but usually do not propose to test their points. It is our contention that questions of meaning in human life (or in societies at large) are too important to be studied only through such largely unreliable means of investigation. What we suggest instead is the scientific investigation of culture, a form of study that approaches matters of meaning in a more accountable way by relying on methods from a scientific repertoire. *Scientific* can be defined here as a kind of reasoning and a kind of research that is based on real evidence, that is, on evidence from the real world, which can be inspected by anyone independently from one's own conviction. [...] What we are proposing, therefore, is a piecemeal contribution to the study of culture, one that bridges the gap between the Humanities and the Natural Sciences, in the realization that both need each other for a better understanding of the world.
>
> Van Peer, Hakemulder and Zyngier (2012: 6-7)

This chapter succinctly describes the methodology used for the two studies reported in this thesis: one relating to the voicing of audio descriptions through TTS (section 2.1.), and another one referring to the machine translation of audio descriptions (section 2.2.). Both studies followed a 2-level structure comprising a pre-test, whose aim was to objectively select either one (i.e. a machine translation engine) or several items (i.e. voices) to be used in the main test, and a main experiment, in which the selected items were assessed in the audio description domain.

Both the pre-test and the main experiment dealing with MT are described in the corresponding articles (chapters 4 and 5), while for TTS only the main experiment is

included in an article (chapter 3). This is why more emphasis will be put on the methodological aspects which could not be made explicit in the articles due to space constraints, that is, the TTS pre-test (see subsection 2.1.1). In this case, results will be reported as they impact directly on the main experiment. Conversely, only key methodological elements will be succinctly explained for the TTS main experiment (subsection 2.1.2), the MT pre-test (subsection 2.2.1) and the MT main experiment (subsection 2.2.2), and no results will be included. This approach has been taken to avoid too much repetition. Table 2.1 gives an overview of where the information related to each test is to be found.

	TTS	MT
Pre-test	Subsection 2.1.1	Subsection 2.2.1 Article 2 (Chapter 4)
Main experiment	Subsection 2.1.2 Article 1 (Chapter 3)	Subsection 2.2.2 Article 3 (Chapter 5)

Table 2.1. Location of the information related to each test

Regarding the global experimental approach, it is worth mentioning that a quasi-experimental within-subjects design was chosen as it was considered to best fit the nature and objectives of our research. The choice of a quasi-experiment was due to the need to estimate the causal impact of manipulating the independent variables (be it voices or raw MT outputs) on the dependent variables, measured by pre-specified indicators (e.g., mean opinion scores or PE effort). A within-subjects design allowed for a higher control over the individual differences of the subjects, drastically reducing the error variance related to them, since all subjects served in all levels of the independent variables.

Regarding the research strategy, using a mixed methods approach (a combination of quantitative and qualitative strategies) allowed us to collect both types of data in the

same experimental session, so that they could complete and in some cases explain each other. However, higher priority was given to quantitative data.

Another relevant aspect of the all studies was the choice of common audiovisual material. In order to obtain results in the period of time scheduled for our research and to optimise the available resources, it was decided that the same audiovisual product from which to extract the stimuli would be used in all cases, and the criteria explained next were followed.

Only foreign-language films were taken into account as they were to be used in the text-to-speech experiment but also in the machine translation study. Various factors were considered in the selection of the language of the films. It is common knowledge that the English language prevails in many sectors of our society. It is the case of the audiovisual production, particularly film production, and of the most state-of-the-art machine translation engines, which unfailingly include English among their main working languages. Moreover, the countries in which audio description is more expanded and largely used are the United Kingdom and the United States, both English-speaking countries. Therefore, English was chosen to be the source language of both the film and the AD.

Since the target language was Catalan, films having also been dubbed and audio described in Catalan were to be selected. At the time the selection took place and putting aside Catalan production films, only dubbed fiction feature films and children animation films had been audio described in Catalan. However, our intended audience were adults, so that children films were disregarded. As for the genre of the films, no particular genre wanted to be favoured since in TTS evaluation studies in other fields the text type was shown to influence the results (Hinterleitner, Neitzel, Möller, & Norrenbrock, 2011). It was therefore decided that the film should belong to a "miscellaneous" category according to Salway, Tomadaki, and Vassiliou's (2004) classification. *Closer* (2004, directed by Mike Nichols) was finally selected due to the prompt availability of all necessary components: the English original script, the English

AD script, the Catalan dubbed script, and the Catalan AD script. The specific selection of the stimuli will be explained when describing each experiment.

As indicated at the beginning of this chapter, the pre-test for the selection of the voices for the TTS experiment will be thoroughly unfolded next, as it has not been described anywhere else. As regards the other experiments, key elements will be succinctly accounted for and only those methodological aspects which were not included in the articles will be explained in more detail.

2.1 Text-to-speech in audio description

TTS AD has already been tested by blind and visually impaired people in other countries (Szarkowska, 2011, Kobayashi et al., 2009) with fairly positive results as far as acceptance is concerned.

According to Huang, Acero, and Hon (2001) and their taxonomy of evaluation methods for TTS systems, both the pre-test and the main experiment in TTS AD could be classified as laboratory human assessments. They are also both analytical and global evaluations at the acoustic-level as they measure general and specific aspects of the voices using human subjects. The fact that experiments were carried out in laboratories allowed for "a higher level of control" over the experiment (Bryman, 2012, p. 55). However, while the pre-test would be a glass-box functional evaluation (Huang et al. 2001) since voices were not assessed in the context of the application, the main experiment would be considered as a black-box judgment assessment for the opposite reason.

2.1.1 Pre-test: Selection of best-rated voices

As previously stated, the aim of this pre-test was to select the voices to be used in the main experiment, in which the reception of automatically and human voiced audio descriptions was to be assessed. Particular emphasis was placed in the objective

approach of the selection, in an effort to leave aside the researcher's subjectivity and own preferences.

In the following paragraphs the methodological aspects taken into account in the design of the test are expounded, justifying each step along the process. First, the voices to be assessed are discussed, followed by the speech samples, the questionnaire design, the sampling population and the test development. The data analysis and the results are finally presented.

2.1.1.1 Voices

To avoid any genre bias, both masculine and feminine voices were included in the test. As for how many voices of each genre should be assessed, two recommendations of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) were followed. First, "[i]t is essential for more than one male and more than one female voice to be used in a balanced design" (ITU-T P.800, 1996, p. 17), and second, "[i]f possible at least five different sources are recommended" (ITU-T P.85, 1994, p. 2). Thus, the number of voices to be assessed amounted to 20: 5 synthetic feminine voices, 5 synthetic masculine voices, 5 natural feminine voices and 5 natural masculine voices.

With regard to synthetic voices, a list of all existing Catalan synthetic voices was made:

- Loquendo: Jordi and Montserrat
- Verbio: Oriol and Meritxell
- Acapela: Laia
- iSpeech: Anna
- Nuance: Núria
- FestCat: 5 masculine and 5 feminine voices
- eSpeak: 1 masculine and 1 feminine voices

eSpeak's voices were dismissed as they were almost unintelligible, and Nuance's feminine voice was not available at the time the selection was made. With respect to FestCat's voices, three masculine voices and one feminine voice were randomly selected to complete the five needed for each genre. The final ones were Jordi and Montserrat by Loquendo, Oriol and Meritxell by Verbio, Laia by Acapela, Anna by iSpeech, and Pep, Jan, Teo and Ona by FestCat.

As for the natural voices, it was determined that the original Catalan AD voice should be ignored for several reasons. First, there was no equivalent "default" synthetic feminine voice. Second, being the "default" voice did not obey any objectively based selection process. Finally, in case it was selected for the main experiment, it would not be available for possible further recordings needed. Instead, only volunteer professional and non-professional voice talents answering a cold calling email (see Annex 2.1. Cold calling email for the recruitment of volunteer voice talents) from the Escola Catalana de Doblatge (ECAD) should be used. The volunteers should have Catalan as their mother tongue (central variant) and present no speech impairments (ITU-T P. 800, 1996).

Both natural and synthetic voices were named A to E for the feminine ones, and F to J for the masculine ones (see Annex 2.3. AD units order for their correspondences)

2.1.1.2 Voice samples

The large number of voices to be assessed (20 in total) led to the decision to use two different speech samples, in order to reduce the fatigue in the respondents and to minimise the learning effect. Thus, one sample was used for the feminine voices and another sample was used for the masculine ones, regardless of whether they were synthetic or natural.

According to ITU-T P.85 (1994), the content of the message to be heard by the participants assessing the voices should belong to the application for which the voices were meant to be used, i.e. the message should be extracted from a film audio

description in our case (see page 47 for an explanation of the selection of the audiovisual product).

In the preparation of the speech samples, many elements were taken into consideration, such as the number, the characteristics and the order of the AD units that would constitute them. In principle, it is encouraged that a large number of sentences constitute the sample, since "[s]ynthesis systems are very inconsistent in that certain combinations of phonemes may engender some audio artefacts while other phoneme or word combinations are extremely smooth and natural" (Viswanathan and Viswanathan, 2005:65). However, a maximum of five sentences (ITU-T P.800, 1996, p. 14) with a maximum duration of 30 seconds per message (ITU-T P.85, 1994, p. 2) is recommended. As a result, it was determined that samples would be composed of 5 AD units each and would not surpass the recommended maximum duration.

In the selection of the AD units, priority was given to the fact that "[i]deally, synthesized sentences played to listeners to be rated must mimic the application domain in which the TTS system is likely to be used – in sentence length and complexity" (Viswanathan & Viswanathan, 2005, p.65). Thus, the high variation in the length of the AD units was deemed characteristic for the AD application domain, and both long and short units were accepted. However, randomising the selection of 5 AD units per each speech sample led to an uneven distribution of AD amount. Since both sample contents needed to be balanced in terms of AD amount, an in depth analysis of the AD script and of the AD units was performed to get to the desired balance taking into account the number of characters of each AD unit and of all AD units together[1].

Thus, the AD units chosen would comprise:

[1] The number of words of the AD units was ignored as the word length factor (Cabeza-Cáceres, 2013: 133) would distort results.

- two of the shortest AD units. Priority was given to the fact that they both had the same number of characters (13 characters) for the sake of balance

- two of the longest AD units (230 and 224 characters each)

- two of the AD units whose length is most frequent (7 AD units 47 characters long)

- two AD units with the average number of characters of all AD units, i.e. 72,4 (71 characters is the closest to the average)

- two 115-character-long AD units, half the number of characters of the longest AD unit, i.e. 230 characters long (116 characters is the closest).

See Annex 2.2. AD unit selection for the final AD unit selection.

The order of the AD units in each sample was balanced, having "no obvious connection of meaning between one sentence and the next" (ITU-T P.800, 1996, p. 14) in any case. Respondents would thus listen 5 times to the same set of AD units presented in 5 different orders, so that fatigue and learning effects would also be reduced (see Annex 2.3. AD units order).

Recording conditions for both samples were as similar as possible and a computer-controlled digital storage system was used. Natural voices were recorded in a professional recording studio (Escola Catalana de Doblatge), while synthetic voices were recorded by the researcher using Sony Vegas Pro 11. Other conditions such as the playing system (Intel Core 2, 4GB RAM, 180GB RAM, floppy drive, CD-ROM, DVD, USB, network connection and Windows XP. Sound card: Realtek ALC262 @ Intel 82801IB ICH9 - High Definition Audio Controller [A-2] PCI) and the listening system (Plantronics Audio 400 DSP), together with the listening environment (multimedia classrooms E and D, Translation and Interpreting Faculty of the UAB), were also controlled, following ITU's recommendation (ITU-T P.800, 1996).

In relation to the listening order of the speech samples, it was established that voice genres would be alternated to diminish the respondents' boredom and that samples

would be counterbalanced to form playlists to cater for the order-of-presentation effect (see Annex 2.4. Listening order of the speech samples).

2.1.1.3 Questionnaire

The International Telecommunication Union defined a testing method based on Mean Opinion Score (MOS) scales for the assessment of the subjective quality of synthetic speech. Many authors have departed from the ITU-T's Recommendation P.85 and proposed modified versions of the initial MOS to create their own evaluation protocols. Our questionnaire was built on the analysis and comparison of several such proposals for text-to-speech human evaluation, including:

- ITU Recommendation P.85 (1994) consisted of 8 items, including seven 5-point scales and one 2-point (yes-no) scale. Items were divided as belonging to the intelligibility factor (listening effort, comprehension problems, articulation), to the quality factor (pronunciation, speaking rate, voice pleasantness) or to both factors at the same time (overall impression, acceptance).
- Viswanathan and Viswanathan's (2005) proposal included 11 items to be assessed on a 5-point scale and divided as belonging to the intelligibility factor (listening effort, pronunciation, comprehension, articulation, speaking rate), to the naturalness factor (naturalness, ease of listening, pleasantness, audio flow) or to both factors at the same time (overall impression, acceptance).
- Cryer, Home, and Morley Wilkins (2010) suggested four types of assessment (functionality, subjective, user testing and technical) in four scenarios (audio book, product, document containing figures, access technology application). The questionnaire relating to the subjective and user testing assessments for audio books was considered to be the most appropriate one taking AD's characteristics into consideration and it included twelve 5-point scales: overall impression, pleasantness, comprehension, pronunciation, prosody, comfortable to listen to for a long period, responsiveness, speaking rate, naturalness, listening effort, and appropriate tone.

- Hinterleitner et al. (2011) focused on the assessment of TTS in audiobook reading tasks and finally proposed eight continuous 7-point rating scales divided as belonging to the prosody factor (intonation, speech pauses, emotion, accentuation), to the listening pleasure factor (voice pleasantness, listening effort, acceptance) or to both factors at the same time (overall impression).

Our final choice included the following items in the following order: general impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort and acceptance. A decision was taken to maintain the first and last items as presented in the other proposals (general impression and acceptance respectively), while the rest of items were ordered from more specific concepts to more general ones: word-focused (accentuation and pronunciation), phrase-focused (speech pauses and intonation), voice-focused (naturalness and pleasantness) and global (listening effort).

A thorough comparison of the way in which questions relating to each item were formulated was performed. Particular attention was paid to Cryer et al.'s (2010) proposal, since it was specifically addressed to blind and partially sighted persons and our questionnaire was to be used at a later stage in the main experiment. In their proposal, up to eight questions kept the same structure ("How would you rate the voice in terms of how…?") while the labels for the answers were the same in seven cases (a. very poor, b. poor, c. acceptable, d. good, e. very good, f. not applicable). This was considered far too monotonous, as respondents should have to listen to the same nine questions with their possible same answers 10 times in one same session (see the explanation on the questionnaire delivery below). Thus, changing the question structure and introducing synonyms was preferred to alleviate the subjects' fatigue and reduce the learning effect. Special care was put in the translation and adaptation of the questions and answer labels from English into Catalan. Two professional translators agreed on the solutions for the translations to sound as natural as possible and to be easily understandable.

The chosen delivery mode was both written and oral, i.e. the questionnaire was designed to be read and heard. Respondents in the pre-test would listen to the pre-recorded questions and possible answers as they would be displayed on the computer screen and would fill in their scores in the prepared Google Forms form (see Annex 2.5. Voice samples' questionnaire) (one form per voice). A decision was taken not to include each item's heading descriptors in the recording, as they were either considered to be too technical, too formal or even difficult to know what exactly they referred to. The order of gradation of the answers was also changed to be ascendant in all cases following Cryer et al. (2010) and Hinterleitner et al. (2011), as it was found to be a more logical progression specially if answers were to be listened to.

2.1.1.4 Participants

For the human assessment of TTS, methods from speech-coding evaluation can be used, in which "*Mean Opinion Score* (MOS) is administered by asking 10 to 30 listeners to rate several sentences of coded speech" (Huang et al., 2001, p. 840). Thus, the sample size was set to 20 subjects. To carefully select the subject population, several inclusion criteria were defined *a priori*. They should be sighted native Catalan speakers, having no hearing impairments and no experience either in TTS or in AD. Controlling for the experience in TTS was deemed important as it has been proven that there is a positive correlation between TTS usage and acceptance: the more people listen to synthetic voices, the more they get used to them and the more they accept them.

Thus, a generic purposive sampling method was chosen based on volunteers answering a cold-calling email (see Annex 2.6. Cold-calling email for the recruitment of volunteer participants) in which inclusion criteria were clearly stated. Subsequently, a snowball sampling method was used, by which the respondents recruited first helped localise other participants with the same characteristics.

Six men and fourteen women were recruited. A heterogeneous group in terms of age (ranging from 19 to 51 years old), education and profession was prioritized in order to avoid any bias in the selection of the voices.

2.1.1.5 Test development

A pilot test with six participants was performed. Genres and age ranges (20-40, 40-60 and over 60) were balanced. They each had different professions and different degrees of education. Three participants listened to synthetic voices, while other three listened to natural voices.

The results obtained in the pilot test showed that some minor changes needed to be performed in the way questions and answers were presented in the forms. They also revealed a lack of complete agreement in what exactly each question referred to. It was decided that a precise explanation of each item would be included in the instructions provided to the participants previous to the test, giving them some examples of what exactly was meant in each case. Some more practical instructions were also added, such as indicating that they would not be allowed to go back to a previous question. Emphasis was also made not to be afraid of giving extreme scores and to avoid comparing the voices, but to try to assess them independently.

The pilot test also allowed for a better definition of the post-questionnaire designed for the collection of the demographics of the participants.

Two experimental sessions were planned due mainly to two reasons. On the one hand, the large number of voices to be assessed would make one only experimental session far too long and exhausting. On the other hand, there is evidence that assessing natural and synthetic speech together has a negative influence in TTS ratings (van Santen, 1993, apud Viswanathan & Viswanathan, 2005, p. 62). Therefore, the first session assessed only synthetic voices, while the second was devoted to natural voices.

In the first session the research was first contextualised and the proceeding of the test was explained in detail, thoroughly describing and illustrating what each question referred to and giving specific instructions on how to fill in the forms (see Annex 2.7. Instructions for the TTS pre-test performance). Participants were then asked to read and sign a participant information sheet and a consent form (see Annex 2.8.

Information sheet and Annex 2.9. Informed consent form), both approved by UAB's Ethical Committee, and to fill in a personal details sheet (see Annex 2.10. Personal details form).

After carrying out a warm-up task with two voices that were not to be used in the actual test and resolving any doubts, participants were asked to assess the 10 synthetic voices, alternating masculine and feminine ones. Participants listened to each voice sample twice with a 5-second pause in between and afterwards. They would then listen to the 9 questions and answer labels pre-recorded by the researcher and would indicate their answers in the form on screen. Once they finished, they would close the form and launch the following voice sample in the order previously established by the researcher for each participant. The test lasted 37 minutes.

In the second session, participants were first asked whether they needed the researcher to remind them of how to proceed. Then the session developed as the first one and natural voices were assessed.

2.1.1.6 Data analysis

The analysis software used to obtain the results was SAS, v9.2 (SAS Institute Inc., USA), fixing the significance level at 0.05.

Descriptive statistics (mean, median, standard deviation, minimum, maximum, lower quartile and upper quartile) were calculated for all items: overall impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort and acceptance.

Overall impression was also analysed using a multinomial model, where the dependent variable was the score of the overall impression and the independent variable was the voice, taking into account that each participant had listened to 5 voices for each case (5 feminine natural, 5 feminine synthetic, 5 masculine natural and 5 masculine synthetic).

If a best rated voice could not be concluded from the overall impression modelling, then the modelling of the acceptance was performed, using the score of the acceptance as the dependent variable and the voice as the independent variable, and taking again into account that each participant had listened to 5 voices for each case (5 feminine natural, 5 feminine synthetic, 5 masculine natural and 5 masculine synthetic).

2.1.1.7 Results

Results were divided up into four groups: feminine natural voices, masculine natural voices, feminine synthetic voices and masculine synthetic voices. Both the pilot test and the experiment results matched.

In relation to the first group, the feminine natural voices, D (professional voice talent) was the one obtaining higher mean and median scores for all items except for the listening effort mean score, for which B obtained a slightly higher mean. The modelling of the overall impression showed that D obtained statistically higher scores than the rest.

As far as the masculine natural voices are concerned, both the overall impression and the acceptance modelling showed no statistically significant differences among F, H and I. However, in this case descriptive statistics helped shed light into which voice was rated higher: H (voice talent student) obtained both the highest median and mean for the overall impression, and the highest mean for the acceptance.

Regarding the feminine synthetic voices, again there were no statistically significant differences between A and C, but A (Laia, by Acapela) presented a higher score in overall impression mean and median, and the highest mean for the acceptance.

Finally, there was no doubt as to which masculine synthetic voice rated highest: H (Oriol, by Verbio) obtained the highest scores in all items and differences were clearly statistically significant.

2.1.2 Main experiment: Reception of text-to-speech audio descriptions

As indicated before, only the basic methodological aspects of the experiment carried out to assess the Catalan TTS audio description and compare their reception to standard human-voiced audio descriptions are included in the section, as the full article is found in chapter 3. First, the voices to be assessed are discussed, followed by the clips, the questionnaires design, the participants and the test development. The results are briefly stated in the end.

2.1.2.1 Voices

The voices selected as best rated in the pre-test were used to voice the experiment's stimuli: a feminine professional voice talent; a masculine voice talent student; Laia, by Acapela, as feminine synthetic voice; and Oriol, by Verbio, as masculine synthetic voice.

2.1.2.2 Clips

As previously explained in the beginning of this chapter, excerpts from the film *Closer*, by Mike Nichols, were selected as stimuli. Two different clips were chosen and randomly assigned a voice gender. Again, careful scrutiny of the AD script was carried out so that two homogeneous and equivalent clips were obtained. A duration of around 3 minutes for the clips was set at the beginning due to the characteristics of the test procedure: in one same session participants should have to watch the clips and answer one questionnaire per clip, which was reckoned to last 23 minutes for the entire session. Longer clips would have increased fatigue in the participants and, therefore, biased the results. Having established the clips' length, neutral content was looked for since the film included some sexually connoted scenes, which could be potentially distracting or offensive. Other variables were also controlled, such as the intervening characters, the background music and the AD density. Table 2.2 presents the main characteristics of both clips (see Annex 3.1. Clips' AD scripts for the AD scripts of the clips):

	Clip 1	**Clip 2**
Length	3 minutes	3 minutes 6 seconds
Intervening characters	Anna and Dan	Anna and Dan
Background music	Opera *Così fan tutte*, by Mozart	Opera *Così fan tutte*, by Mozart
Content	Neutral	Neutral
AD density	571 characters	537 characters

Table 2.2. Clips' main characteristics

The clip recording and their corresponding TTS and human-voiced ADs were performed by a professional sound engineer in the Escola Catalana de Doblatge (ECAD).

2.1.2.3 Questionnaires

The questionnaires designed for the pre-test were also used for the collection of data in the main experiment. However, participants would not read them on screen, but only listen to the questions and possible answers previously recorded by the researcher, who would be in charge of annotating their answers in the corresponding form.

A post-questionnaire (see Annex 3.2. Post-questionnaire) was also prepared to gather information on the participant demographics, preferences and usage of audio described audiovisual products and TTS applications.

2.1.2.4 Participants

According to the International Classification of Diseases (as cited in World Health Organisation Media Centre, 2014), there are four levels of visual function: normal vision, moderate visual impairment, severe visual impairment and blindness. Moderate

and severe visual impairment constitute what is known as “low vision”. All visual impairment is thus represented by low vision together with blindness.

Following this classification, our intended population was blind persons and persons with low vision, i.e. all visually impaired persons in Catalonia[2]. Accordingly, the official statistics website of Catalonia of the Statistical Institute of Catalonia (Idescat) was consulted to know the number of persons legally recognised as being visually disabled.

However, mapping the population was not possible, and neither was creating a sample frame and taking a random sample out of it. Therefore, an a priori purposive sampling approach was employed, followed by a snowballing sampling strategy. Thus, several “criteria for selecting participants [were] established at the outset of the research” (Bryman, 2012, p. 418).

The first criterion to be met was that they should be visually impaired adults, but no other age limitation was established. Actually, an in-depth study of the distribution by age and genre of the persons legally recognised as being visually disabled was carried out. Table 2.3 gives an overview of such distribution for the year 2011:

[2] We will treat “blind persons and persons with low vision”, “blind and partially sighted” and “blind and visually impaired” as synonyms across the thesis.

Age group	Women	Men	Total
18 - 19	44	64	108
20 - 34	628	806	1434
35 - 44	1175	1394	2569
45 - 54	2028	2041	4069
55 - 64	2951	2755	5706
65 - 74	3673	2857	6530
75 and more	6986	3854	10840
Total	**17485**	**13771**	**31256**

Table 2.3. Legally visually disabled population in Catalonia in 2011 by age and genre

A sampling plan was designed (see Annex 3.3. Sampling plan) and an attempt to respect as far as possible the real population distribution was made. However, the desirable number of participants for the age range of 75 and more was not met due mainly to mobility problems related to age.

The second criterion was related to Catalan language knowledge. They should either be native Catalan speakers or, being native Spanish speakers, have bilingual Spanish - Catalan knowledge.

In order to recruit volunteers for the experiment, several blind associations in Catalonia were contacted by email (see

Annex 3.4. Cold-calling email for the recruitment of volunteer participants), such as the Associació Discapacitat Visual Catalunya (ADVC), the Associació Catalana per a la Integració del Cec (ACIC) and the Organización Nacional de Ciegos Españoles (ONCE). Other entities and organisations from UAB were also contacted, including the Observatori per a la Igualtat, the Pograma per a la Integració dels i les Universitàries amb Necessitats Especials (PIUNE) of the Fundació Autònoma Solidaria (FAS), and the Associació Prodisminuïts (ADUAB).

In the end, 67 persons participated in the experiment (see subsection 2.1.2.4 for more details on the participants).

2.1.2.5 Test development

A pilot test with 2 male blind participants in their 50s with different professions was carried out to validate the design with blind persons. No changes needed to be made.

The protocol of the experiment started by the researcher reading an introductory sheet based on the one designed for the pre-test. In this case, the introduction involved the reading of the participant information sheet and consent form by the researcher (see Annex 3.5. Information sheet and Annex 3.6. Informed consent form respectively) and the informed recording of the participants' approval, following the University Ethical Committee's guidelines. After a warm-up task, doubts were resolved and a brief contextualisation of the clips was given. Then, the main experiment started.

A design with counterbalanced measures was used to counteract the order-of-presentation effect. Thus, clips with TTS AD would be presented first, followed by human-voiced AD clips, alternating genders. Orders would be randomised across participants.

Finally, the post-questionnaire was administered (see Annex 3.2. Post-questionnaire).

2.2 Machine translation in audio description

To carry out our main experiment, an MT engine was needed. Many MT engines were freely available for the English-Catalan linguistic combination, but none of them had so far been tested for their performance in the audio description domain. Therefore, our first aim was to choose the most suitable MT engine in the before mentioned linguistic combination and domain, placing special emphasis on the methodology adopted and the evaluation measures used. This was perfomed in what has been called the MT pre-test (see 2.2.1). Once the best performing engine was selected, AD creation, translation and post-editing efforts were compared using both objective and subjective measures (see 2.2.2).

2.2.1 Pre-test: Selection of a machine translation engine

This pre-test aimed to select the MT engine to be used in the main experiment avoiding the researcher's subjectivity in the selection and using both human judgements and automatic metrics. Article 2, included in chapter 4, offers a detailed account of the methodological considerations of the pre-test. However, a general overview of the main aspects will be provided in the next sections. First, the MT engines are presented, followed by the test data, and the evaluation method. Then, a short description of the participants involved and the test procedure is given.

2.2.1.1 MT engines

Free online MT engines for the English-Catalan linguistic combination were sought. Statistically based models (Yandex Translate, by Yandex; Google Translate, by Google; and Bing Translator, by Microsoft) and rule-based models (Apertium, by Universitat d'Alacant; Lucy Kwik Translator, by Lucy Software and Services GmbH) were selected, but no free online hybrid MT model for the required languages could be found to obtain a comprehensive representation of the present MT models.

2.2.1.2 Test data

The AD script of the film *Closer* (see the explanation on the selection of the film in page 47) was exhaustively analysed to find a short neutral clip in terms of content (having no potentially distracting such as sex scenes and/or offensive content) to limit the experiment's duration with the aim of minimising participants' fatigue.

The chosen clip had a duration of 3.09 minutes and comprised 14 different AD units with an AD density of 240 words (1,320 characters) (see Annex 4.1. AD script).

2.2.1.3 MT quality evaluation method

The evaluation model proposed in this pre-test included human judgements, both objective and subjective, and automatic metrics, although emphasis was placed on the first ones. Eight scores were obtained: HBLEU and HTER as automatic objective scores; PE time, as human objective score; and PE necessity, PE difficulty, MT output adequacy, MT output fluency and MT output ranking, as human subjective scores.

The evaluation model included three tasks. The first one was a post-editing task, for which the time was measured, and the HBLEU and the HTER were automatically computed. In this task, the participants were requested to post-edit each AD unit. The second task was an annotation task, in which they were asked to read each translation and rate it by how much they agreed with the statements below:

- "The MT text required no post-editing." This statement related to the PE necessity score, referring to how much post-editing was necessary to fix the translation.
- "The MT text was easy to post-edit." This statement related to the PE difficulty score and referred to how difficult the post-editing necessary to fix the translation was.
- "All the information in the source text was present in the MT text." This statement related to the adequacy score and referred to how much of the meaning expressed in the source was also expressed in the target translation.

- "The MT text is fluent Catalan." This statement related to the fluency score and referred to the extent to which the translation was grammatically well formed and experienced as using natural/intuitive language by a native speaker without taking into account whether the information matches that of the source text or not.

The third task was a ranking task. Participants were presented with a spreadsheet including each original AD unit followed by its five raw MT outputs (see Annex 4.2. Ranking task for the ranking task and Annex 4.3. Raw MT outputs for the raw MT versions). 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".

2.2.1.4 Participants

In order to be able to assess the quality of different translations, the only requirement that participants should meet was to be professional translators in the corresponding linguistic combination, i.e. from English into Catalan. That was the single criterion that was established at the outset of the research for the selection of the participants. No professional audio describers were looked for since audio describers, either in Catalan or in any other language, need not be professional translators as audio description is, in principle, an intersemiotic but intralinguistic activity. Besides, no real skills in AD were needed for the purposes of this test, in which synchronising and adjusting AD units was not required. The main task was the assessment of the quality of 5 different MT systems. Therefore, no other characteristic was necessary.

Five professional, 3 women and 2 men, were directly invited via phone call and participated voluntarily in the test. Their ages ranged from 24 to 45. As far as their education is concerned, all of them had reached at least first-cycle university level studies (bachelor level). In regards to the number of years of experience in the translation field, they ranged from 2.5 to 15. As far as their experience with post-editing is concerned, none of them had professionally worked in the post-editing of machine-translated texts.

2.2.1.5 Test development

The test was designed to be performed in a real-world environment and was calculated to last 4 hours, including the PE tool training. Participants were sent an email (see Annex 4.4. Contextualisation email for the participants) in which definitions of audio description and post-editing were given, the aim of the test stated, the main tasks described, and the items for which they should provide assessments clearly explained. A link to the test-starting document was also added, emphasising the need to avoid interruptions while performing the test. The deadline was a week after the sending of the mail.

The test-starting document (see Annex 4.5. Instructions) included precise instructions for the task and all the links needed for the correct performance of the test, such as the participant information sheet (see Annex 4.6. Information sheet) and consent form (see Annex 4.7. Informed consent form) approved by the University Ethical Committee, the PE tool, the tutorial in both video and written forms for the PE tool (see Annex 4.8. PE tool instructions), the Catalan dubbed version of the clip with the English AD included in the silent gaps as subtitles, the script of both the Catalan dialogues and the English AD in writing (see Annex 4.9. Script of the Catalan dubbed dialogues together with the English AD script), the spreadsheet for the ranking task (see Annex 4.2. Ranking task), and the post-questionnaire form (see Annex 4.10. Post-questionnaire). They were also given a link to the pre-test assessment form (see Annex 4.11. MT pre-test assessment form), which they could optionally fill in.

2.2.2 Main experiment: Comparison of audio description creation, translation and post-editing efforts

The experiment carried out to discern whether implementing MT and post-editing to audio description would actually imply a reduction of the effort involved in the current audio description workflow is unfolded in article 3, included in Chapter 5. The main considerations of the adopted methodology are encapsulated next in separate

sections, including the test data, the effort assessment measures, the participants, the test development and the results.

2.2.2.1 Test data

Since the experiment consisted of three tasks, three neutral clips in terms of content needed to be selected from the film *Closer* (see the explanation on the selection of the film in page 47). The three of them had to be as comparable as possible in terms of duration (for the AD creation task) and AD density (for the AD translation and AD PE tasks). A duration of approximately 3 minutes was set to avoid participants getting too tired in the experimental session, in which they would have to perform three tasks in a row. As for the content, the beginning of the film were disregarded due to specific constraints related to the AD creation of such film part.

Thus, clip A was 3 minutes 9 seconds long and the AD consisted of 14 AD units, 246 words (1,281 characters); clip B was 3 minutes 14 seconds long and the AD consisted of 14 AD units, 212 words (1,147 characters); and clip C was 2 minutes 44 seconds long and the AD consisted of 17 AD units, 271 words (1,401 characters) (see Annex 5.1. Clips' AD scripts).

2.2.2.2 Effort assessment measures and tools

Objective and subjective measures were considered in order to assess the effort each task involved. In relation to objective measures, the key-logging tool InputLog 5.2.01 was used to record and calculate automatically the following elements: total process time and time spent in Subtitle Workshop as indicators of temporal effort; keyboard actions (including total character types and other keystrokes), mouse actions (including clicks, movements and scrolls), switches keyboard to mouse, and total window transitions as indicators of technical effort; and total pause time, mean pause time, number of pauses and pause-to-word ratio as indicators of cognitive effort.

As for the subjective assessments, several questionnaires were used as instruments for the collection of data. A general pre-questionnaire (see Annex 5.2. Pre-questionnaire)

gathered the participants' attitudes to translating and post-editing audio descriptions using a 5-point Likert scale. It also gathered their ratings on a 10-point numerical scale on the effort, creativity impairment, boredom, calque conveyance, and output quality each task under analysis (i.e. AD creation, AD translation, AD PE) would involve.

A general post-questionnaire (see Annex 5.3. Post-questionnaire) gathered the same ratings only changing verb tenses from future to past perfect in the phrasing of the statements to cater for the change from expectations to perceptions, adding a field after each of them to optionally justify their choices.

Three post-task questionnaires (see Annex 5.4. AD creation post-task questionnaire) also included statements to be rated on a 5-point Likert scale, followed by an open field for comments. In the AD translation (see Annex 5.5. AD translation post-task questionnaire) and in the AD PE post-task questionnaires (see Annex 5.6. AD post-editing post-task questionnaire), a question specifically asked whether there were any elements participants had had to adapt from the departure text, such as the amount of information, the length of descriptions, the frequency of descriptions, the number of incomplete sentences (with no verb), or the register (too formal or too colloquial).

2.2.2.3 Participants

Fourteen participants, three male and eleven female, were recruited from among native Catalan-speaking students of an MA in AVT took part in the experiment. All but one had a BA in Translation and Interpreting and all of them finished their MA in Audiovisual Translation in June 2014, when the test took place. They had the same experience as far as AVT and AD creation was concerned.

For technical reasons only the results of twelve could be used. One task of one participant was also lost for technical reasons, but the other two tasks of that same participant could be properly recorded and used.

As for their attitude towards translating and post-editing machine-translated ADs, participants showed a general negative prejudice towards post-editing, while presenting a more positive attitude towards human translation.

2.2.2.4 Test development

The experiment was divided into two parts. In the first part participants were asked to read and sign a participant information sheet and a consent form (see Annex 5.7. Information sheet and Annex 5.8. Informed consent form respectively), both approved by the University Ethical Committee, and to fill in a participant profile form and the general pre-questionnaire. They were then requested to watch the Catalan dubbed version of the film *Closer* from beginning to end uninterruptedly, so that they all had the same contextual information.

After a 30-minute break the second part started, in which they were asked to perform the three tasks in the order they had been assigned. For the AD creation, the audiovisual Catalan dubbed version of the excerpts was provided; for the human translation task, the audiovisual Catalan dubbed version with the audio description in English provided as written text with time codes was given; and for the AD PE task, the audiovisual Catalan dubbed version with the audio description in English, provided as written text with time codes, plus the machine translation generated by Google Translate of the English AD, also provided as written text with time codes, was included. For all of them, specific instructions were also provided.

After each task, they should fill in the corresponding post-task questionnaire. Once tasks were finished, they should fill in the general post-questionnaire.

In this chapter the methodological aspects of the two studies carried out have been expounded, including both the pre-tests and the main experiments which comprise them. More emphasis has been put in the TTS pre-test and its results have been detailed, since this test was the only one which had not been published in the form of

an article. On the contrary, as the MT pre-test and main experiment and the TTS main experiment were included in separate articles, they have been briefly described, but more details can be obtained in the following chapters, which include the corresponding articles.

Chapter 3. Article 1: Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan

3 Article 1

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

Abstract

This article presents an experiment that aims to determine whether blind and visually impaired people would accept the implementation of text-to-speech in the audio description of dubbed feature films in the Catalan context. A user study was conducted with 67 blind and partially sighted people who assessed two synthetic voices when applied to audio description, as compared to two natural voices. All of the voices had been previously selected in a preliminary test. The analysis of the data (both quantitative and qualitative) concludes that most participants accept Catalan text-to-speech audio description as an alternative solution to the standard human-voiced audio description. However, natural voices obtain statistically higher scores than synthetic voices and are still the preferred solution.

Keywords

Accessibility, audio description, audiovisual translation, text-to-speech, speech synthesis, Catalan language, blind, visually impaired

1. Introduction

Accessibility has become a major concern in society in recent decades, and laws are being enforced to guarantee disabled people's rights. Sensorial accessibility should be provided to audiovisual contents: theatres, museums, TV broadcasters and web designers, among others, endeavour to make their contents accessible to persons with disabilities and to comply with regulations.

For users who are blind or visually impaired, audio description (AD) allows them to access visual information (images) appearing on screen, which they would otherwise miss. Audio description can be defined as an inter-semiotic translation in which the visuals are transferred into words that are received aurally by end users (Orero 2007, Orero and Matamala 2007). In films these oral descriptions are inserted in the silent gaps in the dialogue, i.e. when characters are not talking, and create a coherent whole with the film dialogues and soundtrack (Braun 2011). However, because creating and voicing an audio description is a time-consuming and costly process, this access service is not as widely available as one might expect. This is especially striking in social media environments, but also in other traditional broadcasting contexts.

In view of the need for a wider availability of audio described audiovisual products, research on technological processes which fully or partially automate the audio description work flow are considered relevant, from a scientific, social and economic point of view. Within this general framework, this article aims to present the results of a research in which Catalan audio description using text-to-speech (TTS) software was assessed, and compared to standard human-voiced audio descriptions. Our final aim was to find out whether TTS AD in Catalan would be accepted by blind and visually impaired patrons as an alternative solution and to compare the scores attributed to both natural and artificial voices on key aspects. The project's novelty lies in the language under analysis (Catalan). In addition, the methodological approach is also new in comparison with existing text-to-speech audio description tests: on the one hand, it provides a detailed analysis of many features instead of asking about general opinions or perceptions; secondly, it assesses text-to-speech audio description against human-voiced audio description instead of evaluating it in isolation, as further explained in section 3.2.

The article presents, first of all, a review of related work, focussing on text-to-speech audio description but also widening the scope to present other text-to-speech applications in audiovisual translation and media accessibility (section 2).

Methodological aspects are detailed in section 3, and results are discussed in section 4. Conclusions and possibilities for further research close the article.

2. Text-to-speech audio description: an overview

Blind and visually impaired people use text-to-speech in many contexts, and its usefulness has already been proved in different domains. Cryer and Home (2008) analyse the use of synthetic speech technology by blind and partially sighted people. Inspired by Freitas and Kouroupetroglou (2008), they list the many areas in which speech technologies can be used: mobility aids (for instance, GPS navigation devices), educational tools (talking dictionaries, audio textbooks), entertainment (audio subtitles (AST), speaking electronic programming guides) and communication (screen reading software on computers). Speech synthesis seems to offer quicker access to information (Llisterri et al. 1993) and guarantees independence of the user (González García 2004), among other aspects. Cryer and Home (2008) point out two relevant research results of their overview of text-to-speech usage by blind and partially sighted people: firstly, the direct impact of each user experience on the acceptance of synthetic speech, as people gradually get used to synthetic voices; and, secondly, the impact of the naturalness and the context where the artificial voice is being used.

As for text-to-speech audio description (TTS AD), it has been researched within a project developed at the University of Warsaw, Poland, aiming to assess its feasibility and its reception among visually impaired people. Szarkowska (2011: 144) states that "instead of recording a human voice reading out the AD script, TTS AD can be read by speech synthesis software". This guarantees the cost-effectiveness of the AD production in comparison with traditional methods of AD production.

The project analysed the application of TTS AD in several types of audiovisual products:

- in a monolingual feature film in Polish (Szarkowska 2011), where the artificial voice tested was Ewa (female voice), by Ivo Software;
- in a dubbed educational TV series for children (Walczak and Szarkowska 2010), where the artificial voice tested was Ewa (female voice), by Ivo Software;

- in a foreign fiction film, with voice-over (Szarkowska and Jankowska 2012), where the artificial voice tested was Krzysztof (male), by Loquendo;
- in a non-fiction film, with audio subtitling (Mączyńska 2011), where the artificial voices tested were Zosia (female voice) for the AD, and Krzysztof (male voice) for the AST, both by Loquendo;
- in a dubbed feature film (Drożdż-Kubik 2011), where the artificial voice tested was Ewa (female voice), by Ivo Software.

The number of participants ranged from 17 in Drożdż-Kubik (2011) to 76 in Walczak and Szarkowska (2010). The conclusions for each study were as follows: Szarkowska (2011) and Szarkowska and Jankowska (2012) stated that most respondents accepted TTS AD both as an interim solution and as a permanent option, Walczak and Szarkowska (2010) emphasised that most participants enjoyed the voice used in the test, and Mączyńska (2011) and Drożdż-Kubik (2011) explained that a majority of respondents found TTS AD acceptable, although it was not the preferred solution. Hence all five studies showed that most viewers accept TTS in AD.

On a similar note, and inspired by Chapdelaine and Gagnon's work (2009) on an accessible website platform for rendering different levels of audio description (as far as quantity and quality of AD is concerned) on demand, Kobayashi et al. (2009: 249) describe a “technique to use synthesized speech to add ADs to online videos on any websites”. The three steps of their project include determining whether or not synthesized voice can compete with real voices, designing a text-based format to describe the AD scripts, and developing authoring software. Step one is thoroughly explained in Kobayashi et al. (2010): 115 visually-impaired adult participants took part in an informal survey in Japan where three kinds of voice were tested (human, standard TTS, and prototype TTS). This first experiment was followed by an in-depth interview session with three participants. The study continued in the US, where 236 participants completed a survey, followed by an additional in-depth interview session with 8 participants. A follow-up study with 24 participants closed the research. It

included additional variables such as long vs short stimuli, expressive TTS technology vs standard TTS, expert vs novice descriptions, and standard vs extended descriptions. All in all, this broad study showed that synthesized descriptions are generally accepted, especially for relatively short videos and informational content.

With their more experimental approach, Encelle et al. (2011) present an exploratory work on video accessibility for the blind and visually impaired with "audio enrichments composed of speech synthesis and earcons (i.e. nonverbal audio messages" (123). Their study with 21 blind volunteers show that earcons associated with speech synthesis are useful for understanding set-related information, i.e. enriching videos with the use of earcons to complement speech synthesis helps convey visual information.

Moving from academia to industry, the firm Swiss TXT is already planning to offer audio descriptions in which text-to-speech technologies are implemented (Caruso 2012). A web-based editor for transforming text into speech which can be used for audio descriptions has also been developed by Mieskes and Martínez (2011). The editor contains features which allow the speaking rate and pitch to be set, as well as phonetic tuning functionalities. The described scenario would allow a user to upload an existing description or create a new one, upload the corresponding movie and synthesise the descriptions. Similarly, Oncins et al. (2013) have developed the Universal Accessibility System, a multi-language and multi-system mobile application to make live performing arts accessible. The system is designed to offer automatic AD through speech synthesis as well as other features (subtitling, spoken subtitles, an emergency pack, etc.).

Research on text-to-speech in audiovisual translation (AVT) goes beyond audio description and is especially relevant in a strongly related transfer mode: audio subtitling or spoken subtitles, where a synthetic voice is used to automatically read aloud the subtitles and make them accessible not only to blind and visually impaired people, but also to people with reading difficulties. This service has been implemented

in television broadcasts in countries such as the Netherlands (Verboom et al. 2002) and Sweden (De Jong 2006), where two digital boxes are needed to make it work. To expand the availability of spoken subtitles and avoid the need for a special decoder, a user-based device for reading aloud subtitles (Subpal) has been proposed by Nielsen and Bothe (2007), and a free and open-source tool has been developed by Ljunglöf, Derbring and Olsson (2012) within the SubTTS project (Derbring, Ljunglöf and Olsson 2010).

Finally, it is worth mentioning that, focusing exclusively on the language under analysis in this research, Alías, Iriondo and Socoró (2011) present the state of speech synthesis implementation in Catalonia which includes the most relevant companies, research centres and products relating to Catalan synthetic voice generation, and carry out field work to map the actual usage of text-to-speech in Catalan audiovisual media. In a specific section of their article devoted to blind and visually impaired users, they point out that most of them think text-to-speech could be used in AD as long as more natural and expressive voices can be developed, although no specific quantitative data are given.

3. Methodological aspects: materials and method

This section describes the participants involved in the current experiment, the voices used, the film and clip selection process, the evaluation questionnaires drafted, the actual development of the test, and the statistical methods used.

3.1. Participants

Since it was "impossible to map 'the population' from which a random sample" was to be taken (Bryman 2012: 416), an a priori generic purposive sampling strategy was adopted. Such a strategy implied the establishment of certain criteria for selecting participants at the outset of the research. A total of 67 persons participated in the test (55% female, 45% male). The mean age was 52, ranging from 21 years old to 85 years old. 33 participants (49%) were 50 or younger, the others being older (51%). A more detailed distribution of the participants is shown in Table 3.1:

	Women	Men	Total
From 20 to 34	5	4	9 (13%)
From 35 to 44	5	7	12 (18%)
From 45 to 54	7	8	15 (22%)
From 55 to 64	9	6	15 (22%)
From 65 to 74	9	4	13 (19%)
More than 80	2	1	3 (4%)
TOTAL	37 (55%)	30 (45%)	67

Table 3.1. Participants distribution based on sex and age

The age range was not limited to account for the whole spectrum of the adult population to which ADs are offered. Additionally, acceptance of synthetic voices is often linked to their usage, and limiting the age range to younger or older participants would have probably had an effect on the results.

Using the World Health Organisation's classification of visual impairments (2013), 51% of the participants described their disability as blindness, whereas 49% declared it to be low vision, with visual impairment being from birth in 30 cases (45%).

As far as the participants' educational background is concerned, 51% reached at least first degree university level (Bachelor's degree or equivalent), whilst 24% did not reach the first stage of secondary school. 46% reported being unemployed, whilst 13% declared to be employed in clerical posts.

3.2. Voice selection

It was decided that a total of four voices (a male and a female artificial and a male and a female natural voice) would be included in the experiment to avoid any gender bias. In order to select them, a pretest with 20 participants was carried out, as described in Matamala, Fernández-Torné and Ortiz-Boix (2013). Ten synthetic voices (see Table 3.2) and ten natural voices (both professional and non-professional voice artists selected by the Catalan School of Dubbing ECAD) were assessed by the participants (see Fernández-Torné and Matamala 2013 for further details on the methodological aspects of the pretest).

Female	Male
Laia by Acapela	Jan by FestCat
Anna by iSpeech	Teo by FestCat
Meritxell by Verbio	Oriol by Verbio
Montserrat by Loquendo	Jordi by Loquendo
Ona by FestCAt	Pep by FestCat

Table 3.2. Artificial voices

This pretest allowed us to select the voices to be used in the experiment, namely a professional voice talent (a female natural voice), a non-professional but trained voice talent (a male natural voice), Laia by Acapela (a female synthetic voice), and Oriol by Verbio (a male synthetic voice).

3.3. Film and clip selection

The voices were tested in an audio described film excerpt. Various factors influenced the film selection process: first of all, this experiment is part of a wider project in which other technologies such as machine translation are to be tested in the English-Catalan language pair (Fernández-Torné, Matamala and Ortiz-Boix 2012). Therefore, a film which had already been audio described in Catalan (for the TTS AD experiments) and that had also been audio described in English (for the machine translation tests) was required. A dubbed fiction feature film or a children's animation film were the only options, as these were the only dubbed audiovisual products that were audio described in Catalan at the time the experiment took place. Children animation films were disregarded as our intended target audience were adults; hence a dubbed fiction film had to be selected.

Secondly, defining the specific genre was also considered relevant, since in TTS evaluation studies in other fields such as audiobooks, the text type has been shown to have a significant influence on the results. For instance, Hinterleitner et al. (2011) have proven that seven out of the 11 rating scales used in their study were influenced by

the type of text when assessing the quality of one same synthetic voice. Our final decision was not to favour any particular film genre, and a film belonging to a "miscellaneous" category according to Salway et al.'s (2004) classification was chosen.

Finally, from a more practical point of view, it was considered that the availability of the English original script, the English AD script, the Catalan dubbed script, and the Catalan AD script would speed up the research process, and *Closer* (2004, directed by Mike Nichols) was selected. However, to limit the duration of the experiment, it was decided to carry out the experiment using short clips rather than the whole film, unlike the five studies within the TTS project developed at the University of Warsaw and the Jagiellonian University of Krakow (Szarkowska and Jankowska 2012).

As far as the clip selection was concerned, it was decided that two different clips, one clip for female voices and another one for male voices, would be chosen to minimise fatigue and the impact of a learning effect on the subjects. Additionally, an in-depth analysis of the film, of the AD script and of the individual AD units was performed, in order to obtain two comparable clips in terms of content (neutral in both cases, with no potentially distracting and/or offensive content), length (3 minutes in clip 1 vs 3 minutes and 6 seconds in clip 2), intervening characters (Anna and Dan in both clips), background music (the same opera for both), and AD density (571 characters vs 537 characters respectively). Clips were randomly assigned a voice gender, either masculine or feminine, for the audio description.

3.4. Evaluation questionnaires

For the human assessment of synthetic voices, the Telecommunication Standardization Sector of the International Telecommunication Union's (ITU-T) recommends using a Mean Opinion Score (MOS) test, by which listeners are asked to rate several systems taking into account various items (ITU Recommendation P.85 1994), hence this was our chosen approach. The items to be included in our questionnaire were selected after a thorough comparison of various tests in text-to-speech evaluation. These are:

- ITU Recommendation P.85 (1994), which includes seven 5-point scales and one 2-point (yes-no) scale;
- Viswanathan and Viswanathan (2005), who propose 11 items to be assessed on a 5-point scale;
- Cryer, Home and Morley Wilkins (2010), who suggest twelve 5-point scales; and
- Hinterleitner *et al*. (2011), who put forward an evaluation protocol for the assessment of TTS in audiobook reading tasks, concluding to keep eight scales out of the eleven they tested with a continuous 7-point rating scale.

It was finally decided to limit the number of items and to focus on issues directly linked to the end-user reception rather than on the intelligibility dimension, since intelligibility was taken for granted in the selected voices and was deemed more relevant for system performance testing. The final list of items included in our questionnaire is listed next, in the same order as they were presented to participants when given the instructions. Participants assessed each item on a 5-point scale.

Overall impression: a global score, the general opinion participants have of the voice of the audio description.
Accentuation: this score assesses whether the stress is put on the right syllable.
Pronunciation: it measures to what extent words are correctly uttered according to Catalan phonetics.
Speech pauses: it evaluates whether the voice stops when needed between sentence components and between sentences.
Intonation: it assesses whether the pitch curve accurately represents the sentence type (whether it is a question, an exclamation or a declarative sentence).
Naturalness: in synthetic voices, this item assesses to which extent the voice resembles a human voice; in natural voices, it is related to the degree the human voice is forced and dramatised.
Pleasantness: it conveys to what extent the listener finds the voice pleasant.
Listening effort: it involves subjectively assessing whether listening to the voice for a long period of time would be tiring or tedious.

Acceptance: is used to indicate whether the voice is deemed adequate to voice audio descriptions.

It must be stressed that a careful translation into Catalan of each of the previous items, validated by a professional translator and tested in a pilot test, was carried out. It was also decided that heading descriptors were not to be used in the real test since the choice of an oral delivery mode for the test instead of a written one made the use of headings before a question quite awkward and it actually did not enhance comprehension. Therefore, participants were directly asked the questions and read aloud the 5 possible answers to each question preceded by their corresponding score: from least positive (1) to most positive (5) (see Annex 1 for the back translation into English of the actual Catalan questionnaire).

Regarding the order of the items, the overall impression and acceptance items were kept in the first and last positions respectively following the other tests. A logical order was proposed for the remaining scales, from more specific questions to broader ones: word-centered questions (accentuation and pronunciation), phrase-centered questions (speech pauses and intonation), voice-centered questions (naturalness and pleasantness) and a global question (listening effort).

As well as the questionnaire, a post-questionnaire was included, inspired by the works of Walczak (2010), Mączyńska (2011), Chmiel and Mazur (2012) and Pazos (2012). Its aim was to gather information on the participant demographics and to get more subjective information on personal preferences and usage of audio described audiovisual products and TTS applications in devices and/or computers. As in previous studies in the field (Walczak 2010, Mączyńska 2011), such questions were included in a post-questionnaire rather than in a pre-questionnaire. This decision was motivated by our wish to be as tactful as possible, trying not to ask potentially sensitive questions at the beginning of the test. The post-questionnaire, translated from Catalan into English, is included in Annex 2.

3.5. Procedure

Participants did the experiment on a one to one basis in a sound proof booth, following approved ethical procedures. Listening conditions were controlled: the stimuli were played with a VLC Media Player and presented through professional headphones, Beats mixr by Dr. Dre. All participants were volunteers and listened to all stimuli, following a within-subjects design.

The experimental session was initially tested in a pilot test which was developed as follows: participants were given an overview of the project and the actual experiment, and were required to sign a Participant Information Sheet and Consent Form previously approved by the University Ethical Committee. They were instructed to assess each AD voice independently, and a thorough explanation of the nine items for which they were to give ratings was provided (see previous section). A warm-up task using a voice that was not included in the actual experiment was also carried out.

The main experiment then started, and participants were asked to listen to the four voices, replicating always the same pattern: audio stimulus reproduction, 5-second pause, questions 1 to 9 read aloud by the researcher, oral reply by the participant that was written down by the researcher, and a final 3-second pause. The listening order of the voices was randomised across participants, always presenting the synthetic voices first to avoid a negative impact on the TTS system evaluation, as suggested by van Santen (1993), apud Viswanathan and Viswanathan (2005: 62). This part of the experiment lasted 22 minutes and 36 seconds, and the test finished with the post-questionnaire, which was read aloud by the experimenter, who would again write down the answers in the corresponding form.

3.6. Statistical methods

For the eight considered items (accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, acceptance) descriptive statistics (mean, median, standard deviation, minimum, maximum and percentiles) were calculated. Figures 1 and 2 display the means and the medians for all the items. A

multinomial model was established for each item under analysis as the dependent variable and the type of voice as the independent variable. However, some of the items had very low frequencies in some of the categories, so they were recategorized as a binary outcome (scores 1, 2, and 3 were grouped under the category “low score”, whereas scores 4 and 5 were grouped under the category “high score”). Then logistic regression models were used to assess the probability of obtaining a high score.

Overall impression was also analysed using a multinomial model, taking into account the voice, the gender, the age (categorised in under and above 50 to balance groups) and the disability type as independent variables.

All results were obtained using SAS, v 9.2 (SAS Institute Inc, USA). For the decisions, significance level was fixed at 0.05.

4. Results and discussion

From the mean scores of the items (see Figure 3.1) we notice that the natural male voice obtains higher scores in:

- accentuation (4.761, stdev=0.495),
- acceptance (4.687, stdev=0.583),
- intonation (4.478, stdev=0.682),
- listening effort (4.597, stdev=0.605), and
- speech pauses (4.627, stdev=0.624).

The natural feminine voice, by contrast, obtains higher scores in

- pleasantness (4.373, stdev=0.671),
- naturalness (4.522, stdev=0.725),
- overall impression (4.478, stdev=0.725), and
- pronunciation (4.731, stdev=0.479).

The lowest scores for natural voices are related to the female voice acceptance (3.970, stdev=0.244) and intonation (4.343, stdev=0.708). However, for the purposes of our study, what is especially interesting is not which voice gets higher scores on what items, but to observe that the results for the synthetic voices is close to that of natural voices, and that all the scores of the synthetic voices are above 3.1, reaching 4.313 in the accentuation of the synthetic male voice and 4.284 on the pronunciation of the feminine synthetic voice.

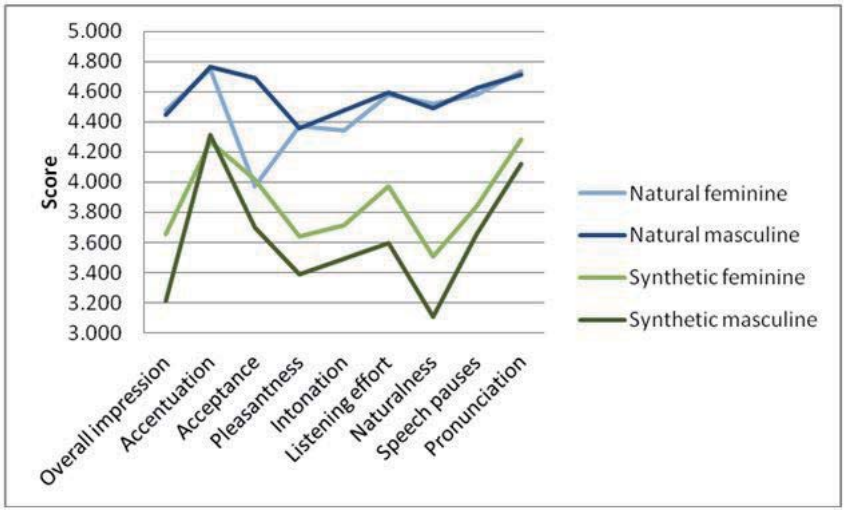


Figure 3.1. Mean scores of all scales for all voices

However, since all items were collected as scores between 1 and 5, the medians (see Figure 3.2) may be more robust than the means. It must be stressed that all median scores are between 3.0 and 5.0. Both male and female natural voices obtain 5.0 in accentuation, listening effort, naturalness, pronunciation, speech pauses, and overall impression, and 4.0 in pleasantness. However, in acceptance and intonation the male natural voice gets higher scores (5.0 vs 4.0). This shows how the more subjective aspects which relate to end users preferences (for instance, acceptance) present greater variation, whilst standard features that a professional describer masters (e.g. accentuation and pronunciation) are more stable. It also shows how even a natural voice may not get the highest mark in terms of pleasantness or intonation.

In as far as artificial voices are concerned, the female voice obtains 4.0 in all items under analysis, whilst the male artificial voice ranges from 3.0 (pleasantness, naturalness, overall impression) to 5.0 (accentuation), with most items rated 4.0 on a 5-point scale (acceptance, intonation, listening effort, speech pauses, pronunciation). Again, what is especially relevant is the fact that all items are assessed above 3.0 and that in some items the median scores are the same for some natural and artificial voices. This is the case of accentuation (same scores for both natural voices and the male artificial voice), acceptance (same scores for the female natural voice and both artificial voices), pleasantness (same scores for natural voices and the female artificial voice), and intonation (same scores for the natural female voice and both artificial voices).

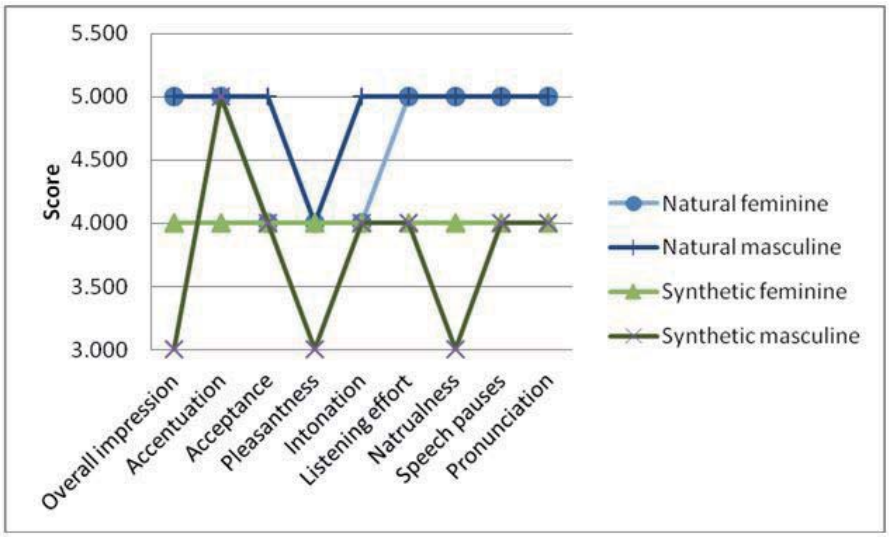


Figure 3.2. Median scores of all scales for all voices

An analysis taking into account the ordinal characteristic of the items is based on the multinomial or logistic models. Statistical significant differences between the synthetic voices and their natural counterparts were found in all items under analysis. In all cases the natural voices were considered to be better than the artificial ones (see Annex 3 for further details).

When comparing the two artificial voices, the analysis shows that the synthetic feminine voice was more accepted, required less effort, was considered to be more natural and obtained a better score in the overall impression than the synthetic masculine one. As for the rest of the items (accentuation, pleasantness, intonation, speech pauses and pronunciation), no statistically significant differences were found.

Focusing on the overall impression, the multinomial model allows us to conclude that women (OR=1.67, IC=(0.96,2.90)) and the group below 50 (OR=1.89, IC=(1.09,3.30)) give statistically significant higher scores than men and people older than 50, respectively. No statistically significant differences were found related to the disability type.

To complement previous statistical analyses, the post-questionnaire provides qualitative data that will be discussed next. When asked about their preferences in terms of a male or a female AD voice, 72% declared they did not have any preferences, with only 16% stating that it depends on the audiovisual product. The reasons for preferring either a female or a male voice in such cases were the topic (in 7 out of the 11 cases, that is 64%) and the characters (in 4 instances, that is 36%).

When asked about their preferences regarding a human or a synthetic voice, 81% of the informants stated that they preferred a human voice to read the AD, 1% declared that they preferred a synthetic voice, 3% said that it depended on the audiovisual product, and 15% declared they did not have any specific preferences as long as the artificial voice sounded natural enough and was not tiring. It must be noted, for example, that in the case of the synthetic voices tested, the naturalness mean scores were 3.507 for the female voice and 3.104 for the male one, and the listening effort mean scores were 3.836 and 3.657 respectively, which are quite strong results in a 5-point scale. It must be also stressed that 51 informants (76%) said they normally use electronic devices with synthetic voice applications on a daily basis.

When explicitly asked about the TTS AD being an alternative solution to human voiced audio description, 94% participants responded positively. Twenty-two participants, i.e.

33%, stated that the main reason for accepting TTS AD as an alternative solution was that it would definitely increase the amount of audio described audiovisual products. Eight out of these 22 participants explained that it would reduce both the costs and time for creating such products. Nine participants (13%) stated that it could be an alternative solution because the quality of synthetic voices was already good enough. On the other hand, 10 informants (15%) stated that synthetically voiced AD was better than no AD at all, while 9 respondents (13%) argued it should only be an alternative, not the usual situation..

When questioned about specific kinds of audiovisual products, the preferences varied slightly, as shown in Figure 3.3: most of the participants agreed on applying TTS AD in documentaries (48 respondents), series (48 respondents) and films (49 respondents); not so many people agreed on applying it to cartoons (36 respondents) and even less informants were willing to implement it in live plays (24 respondents), with 4 participants being against implementing it at all.

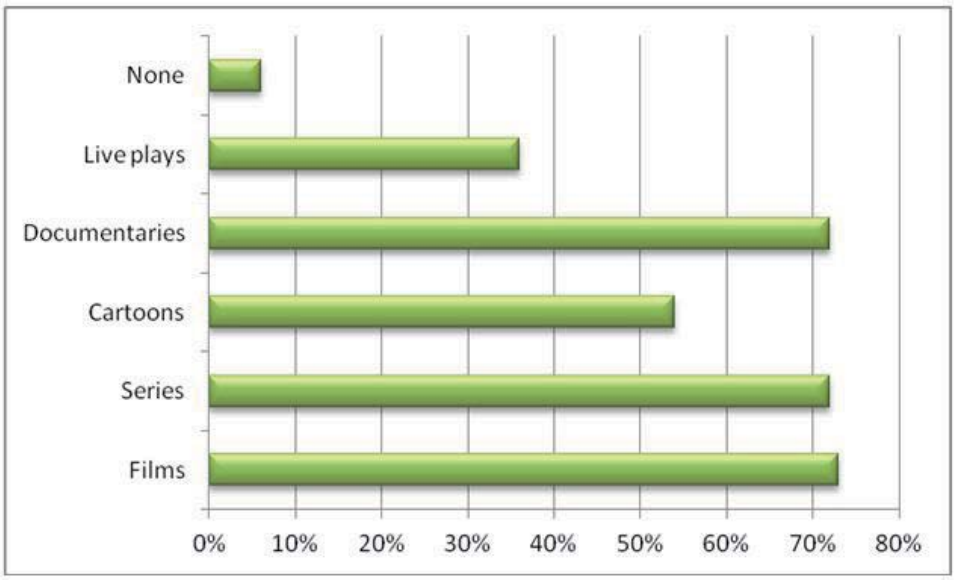


Figure 3.3. Audiovisual products that could be used with TTS AD

Finally, a question about their opinion after listening to the four voices included in the experiment showed a preference for the masculine natural voice (42%) and the feminine natural voice (38%), although 14% said they prefer the feminine synthetic

voice and 6% selected the male synthetic voice. These qualitative data match with the results obtained both in the descriptive and inferential statistics, which actually graded voices in the same order: the natural masculine voice was the one which obtained better mean scores, closely followed by the natural feminine, then the synthetic feminine and finally the synthetic masculine.

5. Conclusions

This article has presented a first analysis of text-to-speech audio description in Catalan, as compared to human-voiced audio descriptions, using both male and female voices. Participants have assessed the voices taking into account various items (overall impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, and acceptance), providing data of both a quantitative and a qualitative nature.

Results show that natural voices in our experiment have statistically higher scores than synthetic voices. They also show that the synthetic feminine voice has higher mean scores than the synthetic masculine voice in all items but accentuation. This proves that the preferential choice of blind and partially sighted persons is the audio description voiced by a human, rather than by a speech synthesis system. This does not mean, though, that TTS AD is not accepted by end users, as shown by the fact that 94% of the participants consider TTS an alternative acceptable solution, and 20% of the respondents actually state that their preferred voice from the four under analysis is a synthetic one. Moreover, it is particularly relevant that no mean score of any of the items goes under 3.1 on a 5-point scale. As an example, the acceptance item’s lowest score is a 3.7 (for the synthetic masculine voice) and the overall impression item's lowest score is a 3.2 (also for the synthetic masculine voice).

This experiment follows previous research on TTS AD carried out in Poland and Japan but it is the first of its kind in Catalan. However, it also has its own limitations. First of all, since the study used a non-probability sampling approach (Bryman 2012:418), the results cannot be generalised to the whole Catalan blind and partially sighted

population. Another of its setbacks is the length of the clips: it remains to be seen whether the results would remain the same in longer productions and in various genres. It would also be highly interesting to see whether reception varies in productions originally shot in Catalan and in dubbed productions, since the language and the sound conditions are different. Another topic worth researching would not only be the perceived quality based on a list of previously selected items, but also the engagement of the audience, in line with Fryer and Freeman's research (2013b). Finally, it would also be worth researching end users behaviour if given the possibility of tuning their own AD preferences, at least as far as voice, voice gender and volume are concerned, in line with Walczak and Szarkowska's approach (2012).

All in all, it is our hope that this type of research will allow us to find new ways of increasing access to culture and entertainment for the blind and visually impaired, both on traditional and new media. We are convinced that speech technologies but also other language and visual processing technologies will play a key role and will open a myriad of research possibilities.

Acknowledgements

This work has been carried out within the scope of the doctoral program in Translation and Intercultural Studies offered in the Department of Translation and Interpreting at the Universitat Autònoma de Barcelona, with the financial support of the projects TECNACC "Technologies for accessibility" (APOSTA 2011-010) and ALST ("Linguistic and sensorial accessibility", project ref. code FFI2012-31024 of the Spanish Ministerio de Ciencia e Innovación). Anna Matamala is a TransMedia Catalonia member, funded by Generalitat de Catalunya (2014SGR27). We would like to thank Iola Ledesma and Jose Navarro, of the Escola Catalana de Doblatge (ECAD), for helping in the selection of the natural voices and in the recording of the clips used in the experiments, as well as Ana Vázquez and Anna Espinal, of the Applied Statistics Service of the Autonomous University of Barcelona, for their collaboration in defining the sample and in the

statistical analysis of the data. Special thanks to the Associació Discapacitat Visual de Catalunya (ADVC) and Associació Catalana per a la Integració del Cec (ACIC) for their kindness and support and for providing us with so many volunteers for the experiment.

Bibliography

Alías, Francesc, Iriondo, Ignasi and Joan Claudi Socoró (2011). "Aplicació de tècniques de generació automàtica de la parla en producció audiovisual". *Quaderns del CAC*, 37(1), 105-114. http://www.cac.cat/pfw_files/cma/recerca/quaderns_cac/Q37_Alias_etal.pdf (consulted 30.07.2014)

Braun, Sabine (2011). "Creating coherence in Audio Description." *Meta: Journal des Traducteurs / Meta: Translator's Journal*, 56(3), 645-662. http://www.erudit.org/revue/meta/2011/v56/n3/1008338ar.pdf (consulted 30.07.2014)

Bryman, Alan (2012). *Social Research Methods*. Oxford, UK: Oxford University Press.

Caruso, Beatrice (2012) "Audio Description Using Speech Synthesis". Languages and the Media. 9th International Conference on Langauge Transfer in Audiovisual Media. Conference Catalogue. Berlin, Germany: ICWE, 59-60.

Chapdelaine, Claude and Langis Gagnon (2009). "Accessible Videodescription On-Demand". *ASSETS '09. Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. New York, USA: ACM, 221-222.

Chmiel, Agnieska and Iwona Mazur (2012). "AD reception research: Some methodological considerations". Elisa Perego (ed.) (2012). *Emerging topics in translation: Audio description*. Trieste, Italy: EUT Edizioni Università di Trieste, 57-80.

Cryer, Heather and Sarah Home (2008). *Exploring the use of synthetic speech by blind and partially sighted people*. Literature review #2. Birmingham: RNIB Centre for Accessible Information (CAI).

Cryer, Heather, Home, Sarah, and Sarah Morley Wilkins (2010). *Synthetic speech evaluation protocol*. Technical report #7. Birmingham: RNIB Centre for Accessible Information (CAI).

De Jong, Frans (2006). "Access Services for Digital Television". Ricardo Perez-Amat and Álvaro Pérez-Ugena (eds) (2006). *Sociedad, integración y televisión en España*. Madrid, Spain: Laberinto Comunicación, 331-344.

Derbring, Sandra, Ljunglöf, Peter and Maria Olsson (2009). "SubTTS: Light-weight automatic reading of subtitles". Kristiina Jokinen and Eckhard Bick (eds) (2009). *Nodalida'09: Proceedings of the 17th Nordic Conference of Computational Linguistics*. NEALT Proceedings Series, vol. 4. Odense, Denmark: Northern European Association for Language Technology.

Drożdż-Kubik, Justyna (2011). "Harry Potter i Kamień Filozoficzny słowem malowany – czyli badanie odbioru filmu z audiodeskrypcją z syntezą mowy". MA Thesis. Jagiellonian University.

Encelle, Benoît, Ollagnier-Beldame, Magali, Pouchot, Stéphanie and Yannick Prié (2011). "Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis". *ASSETS '11: Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 123-130.

Fernández-Torné, Anna, Matamala, Anna and Carla Ortiz-Boix (2012). "Technology for accessibility in multilingual settings: the way forward in AD?" Paper presented at *The translation and reception of multilingual films* (University of Montpellier 3, 15-16 June 2012). http://ddd.uab.cat/record/117160 (consulted 30.07.2014)

Fernández-Torné, Anna and Anna Matamala (2013). "Methodological considerations for the evaluation of TTS AD's acceptance in the Catalan context". Paper presented at *ARSAD (Advanced Research Seminar).* (Autonomous University of Barcelona, 13-14 March 2013). http://ddd.uab.cat/record/117078 (consulted 30.07.2014)

Freitas, Diamantino and Georgios Kouroupetroglou (2008). "Speech technologies for blind and low vision persons". *Technology and Disability* 20, 135-156.

Fryer, Louise and Jonathan Freeman (2013). "Visual impairment and presence: measuring the effect of audio description". *Proceedings of the 2013 Inputs-Outputs Conference: An Interdisciplinary Conference on Engagement in HCI and Performance*. New York, USA: ACM, article no. 4.

González García, L (2004). "Assessment of text reading comprehension by Spanish-speaking blind persons". *British Journal of Visual Impairment* 22 (1), 4-12.

Hinterleitner, Florian et al. (2011). "An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks". *Proceedings of the Blizzard Challenge Workshop*. International Speech Communication Association (ISCA).

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*. Geneva, Switzerland: ITU. http: //www.itu.int/rec/T-REC-P.85-199406-I/en (consulted 23.07.2014)

Kobayashi, Masatomo, Fukuda, Kentarou, Takagi, Hironobu and Chieko Asakawa (2009). "Providing synthesized audio description for online videos". *ASSETS '09: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 249-250.

Kobayashi, Masatomo, O'Connell, Trisha, Gould, Bryan, Takagi, Hironobu and Chieko Asakawa (2010). "Are Synthesized Video Descriptions Acceptable?" *ASSETS '10: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, USA: ACM, 163-170.

Ljunglöf, Peter, Derbring, Sandra and Maria Olsson (2012). "A free and open-source tool that reads movie subtitles aloud." *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*. Montreal, Canada: Association for Computational Linguistics (ACL), 1-4.

Llisterri, Joaquim, Fernàndez, Natividad, Gudayol, Francesc, Poyatos, Juan José and Josep Martí (1993). "Testing users' acceptance of Ciber232, a test to speech system used by blind people". Granström, B., Hunnicutt, S., and Spense, K.E. (eds) (1993) *Speech and Language Technology for Disabled Persons. Proceeding of an ESCA Workshop*. Stockholm, Sweden. 203-206.

Mączyńska, Magdalena (2011). *TTS AD with audio subtitling to a non-fiction film. A case study based on* La Soufriere *by Werner Herzog*. MA Thesis. University of Warsaw.

Matamala, Anna, Fernández-Torné, Anna and Carla Ortiz-Boix (2013). "Enhancing sensorial and linguistic accessibility: further developments in the TECNACC and ALST projects". Paper presented at the *5th International Conference Media for All. Audiovisual Translation: Expanding Borders*. (Dubrovnik, Croatia, 25-27 September 2013). http://ddd.uab.cat/record/116868 (consulted 30.07.2014)

Mieskes, Margot and Juan Martínez Pérez (2011). "A Web-based Editor for Audio-titling using Synthetic Speech". *3rd International Symposium on Live Subtitling with Speech Recognition*. Antwerp, Belgium. http://www.respeaking.net/Antwerp%202011/Webbased_editor.pdf (consulted 30.07.2014)

Nielsen, Simon and Hans-Heinrich Bothe (2007). "SUBPAL: A Device for Reading Aloud Subtitles from Television and Cinema". Marion A. Hersh and James Ohene-Djan (eds) (2008). *Proceedings of the Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Assistive Technology for All Ages CVHI 2007*. CEUR Workshop Proceedings, vol. 415. http://ceur-ws.org/Vol-415/paper17.pdf (consulted 30.07.2014)

Oncins, Estel·la, Lopes, Oscar ; Orero, Pilar, Serrano, Javier and Jordi Carrabina (2013). "All together now: a multi-language and multi-system mobile application to make living performing arts accessible". *Jostrans* 20, 147-164.

Orero, Pilar (2007). "Sampling audio description in Europe". Jorge Díaz Cintas, Pilar Orero and Aline Remael (eds) (2007). *Media for All. Subtitling for the Deaf, Audio Description, and Sign Language*. Amsterdam/New York: Rodopi, 111-125.

Orero, Pilar and Anna Matamala (2007). "Accessible opera: overcoming linguistic and sensorial barriers". Perspectives. Studies in Translatology 15(4), 262-277. https://ddd.uab.cat/record/117149 (consulted 30.07.2014)

Pazos, Patricia (2012). *Audiosubtitulación: una posible solución para la accesibilidad a los medios audiovisuals.* MA thesis. Autonomous University of Barcelona.

Salway, Andrew, Tomadaki, Elia and Andrew Vassiliou (2004). *Building and analysing a corpus of AD scripts. TIWO Television in Words. Report on Worckpackage 2*. Surrey, UK: University of Surrey.

Szarkowska, Agnieska (2011). "Text-to-speech audio description: towards wider availability of AD". *The Journal of Specialised Translation* 15, 142-162.

Szarkowska, Agnieska and Anna Jankowska (2012). "Text-to-speech audio description of voice-over films. A case study of audio described *Volver* in Polish". Elisa Perego (ed.) (2012). *Emerging topics in translation: Audio description*. Trieste, Italy: EUT Edizioni Università di Trieste, 81-98.

Verboom, Maarten, Crombie, David, Dijk, Evelien and Mildred Theunisz (2002). "Spoken Subtitles: Making Subtitled TV Programmes Accessible". Klaus Miesenberger, Joachim Klaus and Wolfgang L. Zagler (eds) (2002). *Proceedings of Computers Helping People with Special Needs, 8th International Conference, ICCHP 2002.* Berlin-Heidelberg, Germany: Springer-Verlag, 295-302.

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.

Walczak, Agnieska (2010). *Audio description for children. A case study of text-to-speech audio description of educational animation series* Once Upon a Time... Life. MA Thesis. University of Warsaw.

Walczak, Agnieska and Agnieska Szarkowska (2012). "Text-to-speech audio description of educational materials for visually impaired children". Silvia Bruti and Elena Di Giovanni (eds) (2012). *Audio Visual Translation across Europe: An Ever-Changing Landscape*. Berna/Berlin: Peter Lang, 209-234.

Websites

World Health Organisation (2013) Fact Sheet no. 282. http://www.who.int/mediacentre/factsheets/fs282/en/ (consulted 17.07.2014)

Annex 1. Questionnaire

How would you describe the quality of the voice you have just heard?

1. Bad
2. Regular
3. Neutral
4. Good
5. Excellent

Did you detect anomalies in terms of the accentuation of words?

1. Yes, al lot of them
2. Yes, many
3. Yes, some
4. Yes, but only a few
5. No, none

Did you notice anomalies in terms of pronunciation?

1. Yes, al lot of them
2. Yes, many
3. Yes, some
4. Yes, but only a few
5. No, none

Do you think the voice makes pauses when it is needed?

1. No, never
2. No, almost never
3. Yes, normally
4. Yes, almost always
5. Yes, always

How would you rate the intonation of sentences?

1. Very bad
2. Bad
3. Good
4. Quite good
5. Very good

How would you define the degree of naturalness of the voice?

1. Very unnatural
2. Unnatural
3. Natural
4. Quite natural
5. Very natural

To what extent do you deem this voice to be pleasant?

1. Very unpleasant
2. Unpleasant
3. Neutral
4. Pleasant
5. Very pleasant

Do you think listening to this voice for a long time would be tiring?

1. Yes, a lot
2. Yes, quite a lot
3. Yes, a little bit
4. No, not much
5. No, not at all

Do you think this voice could be used for voicing audio descriptions?

1. No, never

2. No, almost never

3. Yes, in some cases

4. Yes, in many cases

5. Yes, always

Annex 2. Post-questionnaire

*Mandatory field

- Identifier *

Enter your initials (first name initial, first surname initial and second surname initial) followed by your age. Do not leave any blank space in between.

- Age*

- Sex*

Male / Female

- Level of studies reached*

Lower than first stage of secondary school

Secondary education, first stage

Secondary education, second stage

Advanced vocational education

First cycle university education (diploma, degree, engineering or graduate studies)

Second cycle university education (master, postgraduate or doctoral studies)

- In case you have reached university education, please specify.

- Occupation*

Public administration management and management of companies with 10 or more wage earners.

Management of companies with less than 10 wage earners

Management of companies without wage earners

Professions associated with 2nd and 3rd cycle university degrees and the like

Professions associated with a 1st cycle university degree and the like

Support technicians and professionals

Administrative type employees

Catering services workers and personal services workers

Protection and security service workers

Retail workers and the like

Workers skilled in agriculture and fishing

Skilled construction workers, except machinery operators

Skilled workers in the extractive industry, metallurgy, construction of machinery and related trades.

Skilled workers from the graphic arts, textile and tailoring, elaboration of food, cabinetmakers, craftspersons and other similar industries

Fixed machinery and industrial installation operators; fitters and assemblers.

Mobile machinery drivers and operators

Unskilled workers in the service sector (except transports)

Agriculture, fishing, construction, manufacturing industries and transport labourers.

Armed forces

Unemployed for longer than one year

Unemployed, seeking a first job

- Profession in your own words

- Kind of visual impairment according to WHO*

Blindness / Low vision

- How long have you been visually impaired for? *

From birth / For less than 1 year / For between 1 and 10 years / For between 11 and 20 years / For more than 20 years any

- Have you ever seen an audio described product (films, series, theatre plays, etc.)?*

Yes / No

- In case you have, which kind of products? (You can tick more than one answer)

Films / Series / Cartoons / Theatre plays / Opera plays

- How often do you use audio described products?*

At least once a day / At least once a week / At least once a month / Never / Other

- Do you prefer the AD to be read by*

A man / A woman / It depends on the audiovisual product / I don't care

- If it depends on the audiovisual product, what does it depend on exactly?

- You prefer the AD to be read by*

A human voice / An artificial voice / It depends on the audiovisual product / I don't care

- If it depends on the audiovisual product, what does it depend on exactly?

- Do you use electronic devices with synthetic voice applications, such as mobile phones or computers?*

Yes / No

- How often do you use them?*

At least once a day / At least once a week / At least once a month / Never

- Have you ever used audio described products with synthetic voice?*

Yes / No

- Do you think it is an alternative solution to human voiced audio description?*

Yes / No

- Why do you think so?*

- What kind of products would you use with synthetic voiced AD? (You can tick more than one answer)*

Films / Series / Cartoons / Documentaries / Live plays / None

- Which voice, from the 4 voices you have just heard, did you like the most?*

The masculine synthetic voice / The masculine natural voice / The feminine synthetic voice / The feminine natural voice

- Would you be able to rank them in order, from the one you like the most to the one you liked the least?

- Other comments

Annex 3. Odds ratio (OR) tables

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.3878	0.2041	0.7367
Synthetic masculine vs natural masculine	0.06400	0.03085	0.1328
Synthetic masculine vs natural feminine	0.05216	0.02484	0.1095
Synthetic feminine vs natural masculine	0.1650	0.08252	0.3300
Synthetic feminine vs natural feminine	0.1345	0.06652	0.2719
Natural masculine vs natural feminine	0.8150	0.4144	1.6029

Overall impression

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	1.1856	0.5856	2.4005
Synthetic masculine vs natural masculine	0.1604	0.06801	0.3783
Synthetic masculine vs natural feminine	0.1774	0.07617	0.4129
Synthetic feminine vs natural masculine	0.1353	0.05743	0.3187
Synthetic feminine vs natural feminine	0.1496	0.06433	0.3478
Natural masculine vs natural feminine	1.1058	0.4334	2.8212

Accentuation

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.6593	0.3323	1.3081
Synthetic masculine vs natural masculine	0.1128	0.05074	0.2506
Synthetic masculine vs natural feminine	0.1000	0.04445	0.2251
Synthetic feminine vs natural masculine	0.1710	0.07730	0.3784
Synthetic feminine vs natural feminine	0.1517	0.06774	0.3398
Natural masculine vs natural feminine	0.8871	0.3765	2.0898

Pronunciation

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.6419	0.3409	1.2088
Synthetic masculine vs natural masculine	0.08431	0.04036	0.1761
Synthetic masculine vs natural feminine	0.1045	0.05101	0.2142
Synthetic feminine vs natural masculine	0.1313	0.06367	0.2710
Synthetic feminine vs natural feminine	0.1628	0.08043	0.3297
Natural masculine vs natural feminine	1.2397	0.5845	2.6297

Speech pauses

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.6922	0.3463	1.3835
Synthetic masculine vs natural masculine	0.1176	0.04655	0.2973
Synthetic masculine vs natural feminine	0.1568	0.06638	0.3701
Synthetic feminine vs natural masculine	0.1700	0.06693	0.4316
Synthetic feminine vs natural feminine	0.2265	0.09541	0.5375
Natural masculine vs natural feminine	1.3324	0.4610	3.8508

Intonation

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.4920	0.2652	0.9126
Synthetic masculine vs natural masculine	0.07831	0.03867	0.1586
Synthetic masculine vs natural feminine	0.07904	0.03905	0.1600
Synthetic feminine vs natural masculine	0.1592	0.08044	0.3150
Synthetic feminine vs natural feminine	0.1607	0.08124	0.3177
Natural masculine vs natural feminine	1.0092	0.4981	2.0447

Naturalness

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.5685	0.2815	1.1482
Synthetic masculine vs natural masculine	0.07245	0.02695	0.1948
Synthetic masculine vs natural feminine	0.08628	0.03367	0.2211
Synthetic feminine vs natural masculine	0.1274	0.04734	0.3431
Synthetic feminine vs natural feminine	0.1518	0.05913	0.3895
Natural masculine vs natural feminine	1.1909	0.3705	3.8281

Pleasantness

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.4473	0.2363	0.8468
Synthetic masculine vs natural masculine	0.1148	0.05619	0.2347
Synthetic masculine vs natural feminine	0.1205	0.05921	0.2453
Synthetic feminine vs natural masculine	0.2568	0.1272	0.5181
Synthetic feminine vs natural feminine	0.2695	0.1340	0.5417
Natural masculine vs natural feminine	1.0494	0.5012	2.1973

Listening effort

Voice	OR Estimate	OR IC Lower	OR IC Upper
Sinthetic masculine vs synthetic feminine	0.3832	0.1814	0.8096
Synthetic masculine vs natural masculine	0.07689	0.02486	0.2378
Synthetic masculine vs natural feminine	0.01831	0.002365	0.1417
Synthetic feminine vs natural masculine	0.2007	0.06254	0.6438
Synthetic feminine vs natural feminine	0.04778	0.006047	0.3775
Natural masculine vs natural feminine	0.2381	0.02552	2.2217

Acceptance

Chapter 4. Article 2: Machine Translation Evaluation through Post-Editing Measures in Audio Description

4 Article 2

Fernández-Torné, A. (Forthcoming). Machine Translation Evaluation through Post-Editing Measures in Audio Description. *inTRAlinea*. 2016, 18.

Abstract

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

1. 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.

2. 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 online 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 e*x 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.

3. Experimental Set-Up

Various aspects related to the study design are described next.

3.1. 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[3], 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.

3.2. 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

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

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 **Appendixes**

Table A.1).

3.3. 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[4]:

[4] Search performed in September 2013.

- 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.

3.4. 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 4.1):

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

Table 4.1. 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 4.1, 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'.

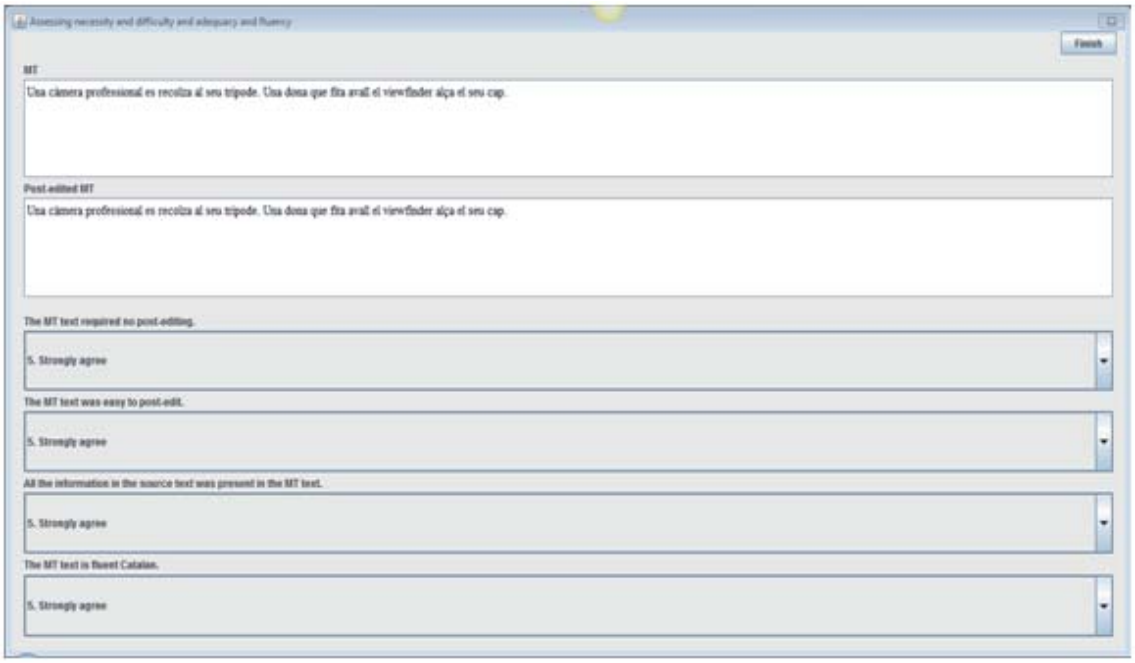
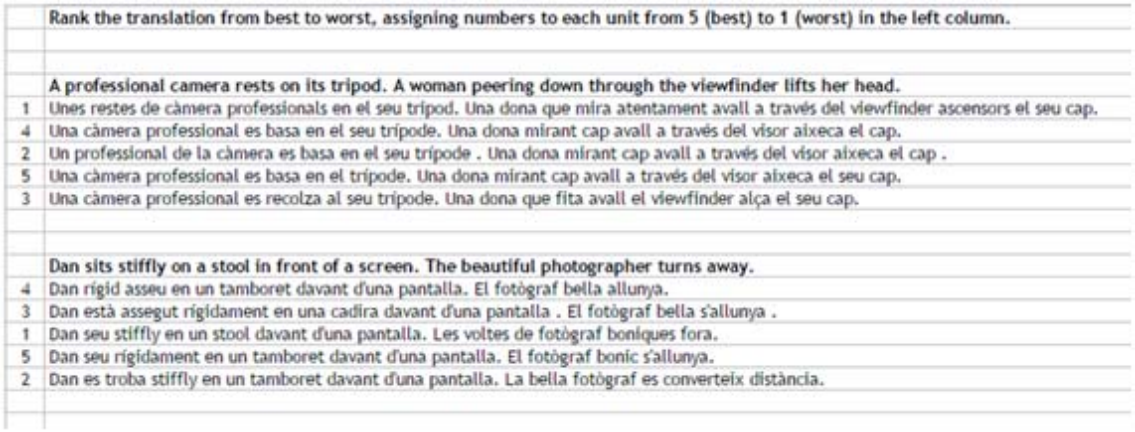


Figure 4.1. Subjective assessments per AD unit

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 4.2:



	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 trípod. 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 .
5	Una càmera professional es basa en el trípode. Una dona mirant cap avall a través del visor aixeca el seu cap.
3	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.
4	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.
5	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 4.2. Ranking of raw MT outputs

3.4. 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, the HTER score, and several edit operations, among others.

3.5. 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[5]. After post-editing each unit, participants were

[5] 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.

asked to provide their evaluations on PE difficulty, PE necessity, MT adequacy and MT fluency, while PE time and HTER were 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.

4. 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.

5. 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.

5.1.HBLEU

HBLEU metrics were obtained using the Language Studio^{TM} Pro Desktop Tools package,

by Asia Online. Table 4.2 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 4.2. 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).

5.2. HTER

HTER metrics were obtained using the Language Studio™ Pro Desktop Tools package, by Asia Online. Table 4.3 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

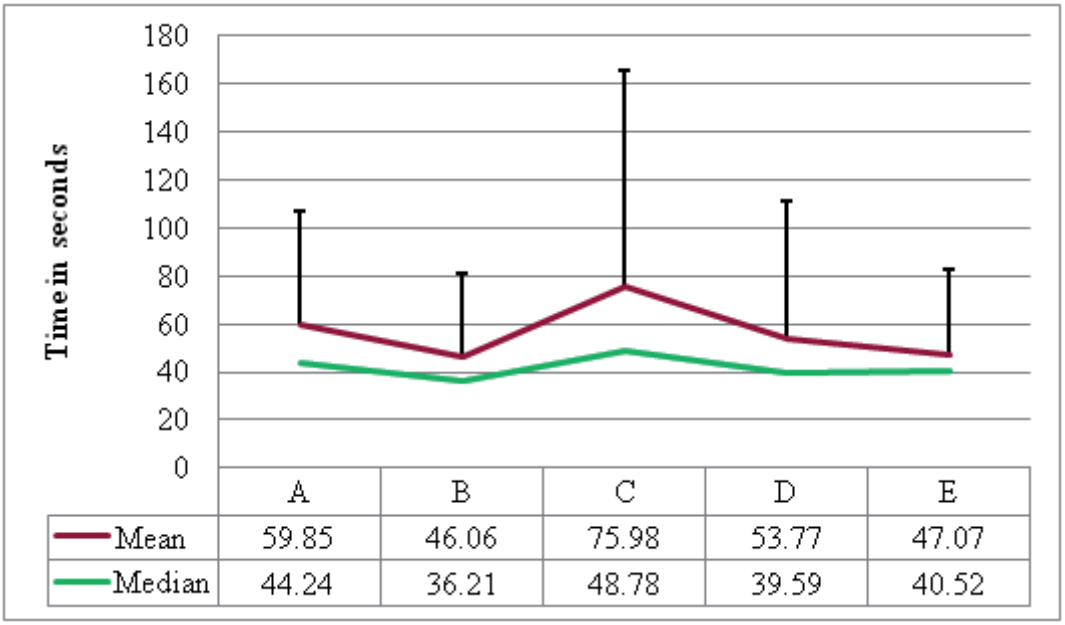
Table 4.3. HTER scores

5.3. 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 4.3). 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.

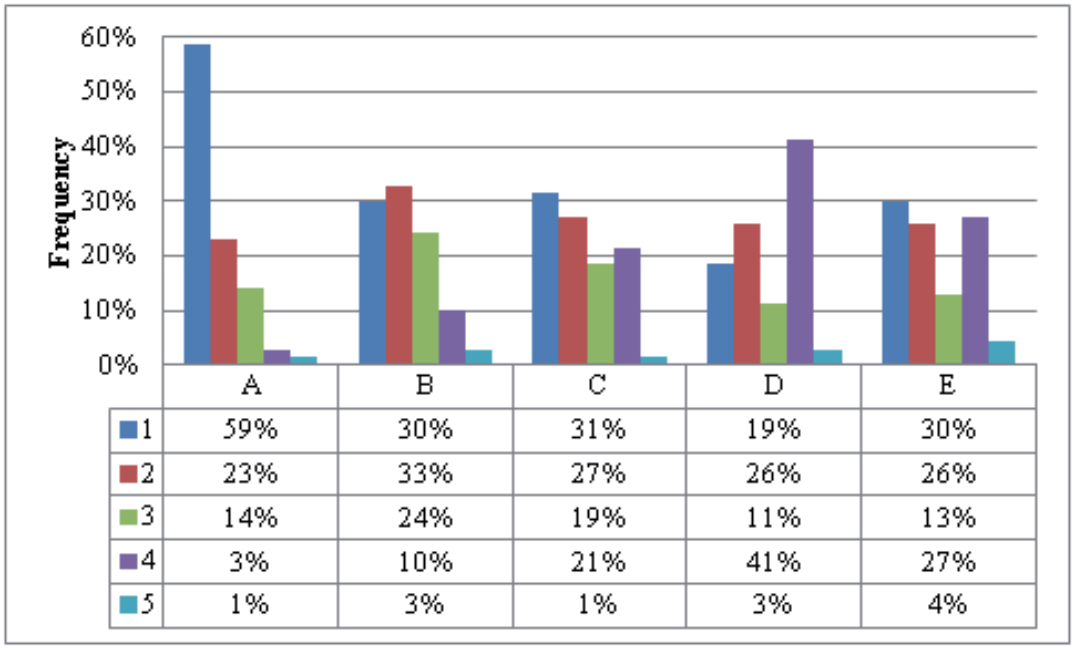


	A	B	C	D	E
Mean	59.85	46.06	75.98	53.77	47.07
Median	44.24	36.21	48.78	39.59	40.52

Figure 4.3. Mean and median PE times per system with standard deviation error bars

5.4. PE necessity

According to Figure 4.4 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).



	A	B	C	D	E
1	59%	30%	31%	19%	30%
2	23%	33%	27%	26%	26%
3	14%	24%	19%	11%	13%
4	3%	10%	21%	41%	27%
5	1%	3%	1%	3%	4%

Figure 4.4. 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).

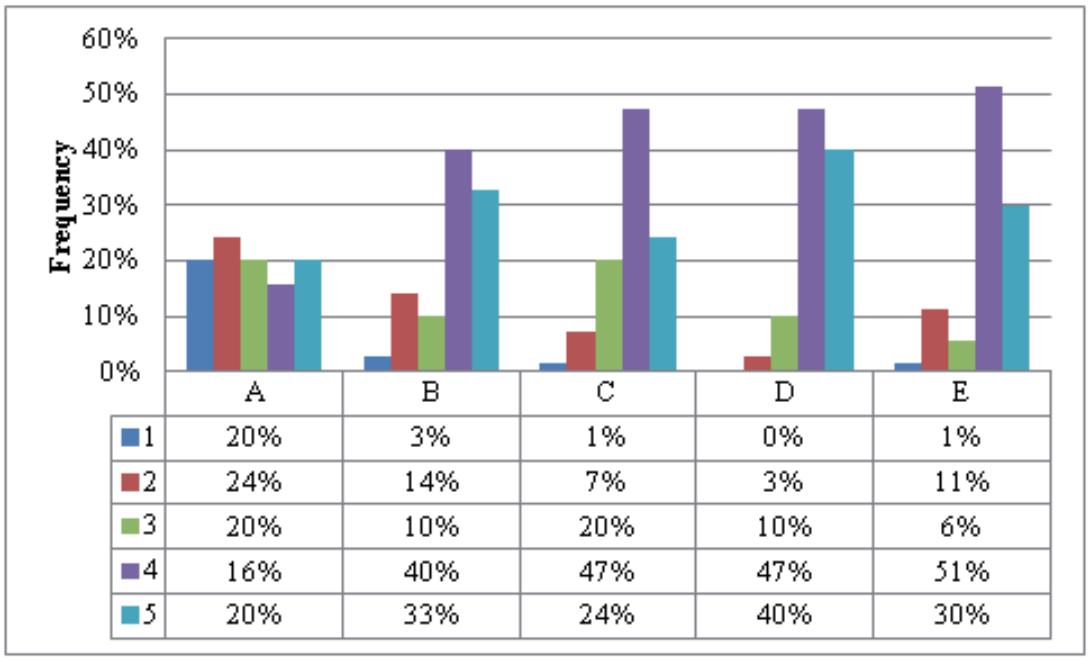
5.5. PE difficulty

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 4.5 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%

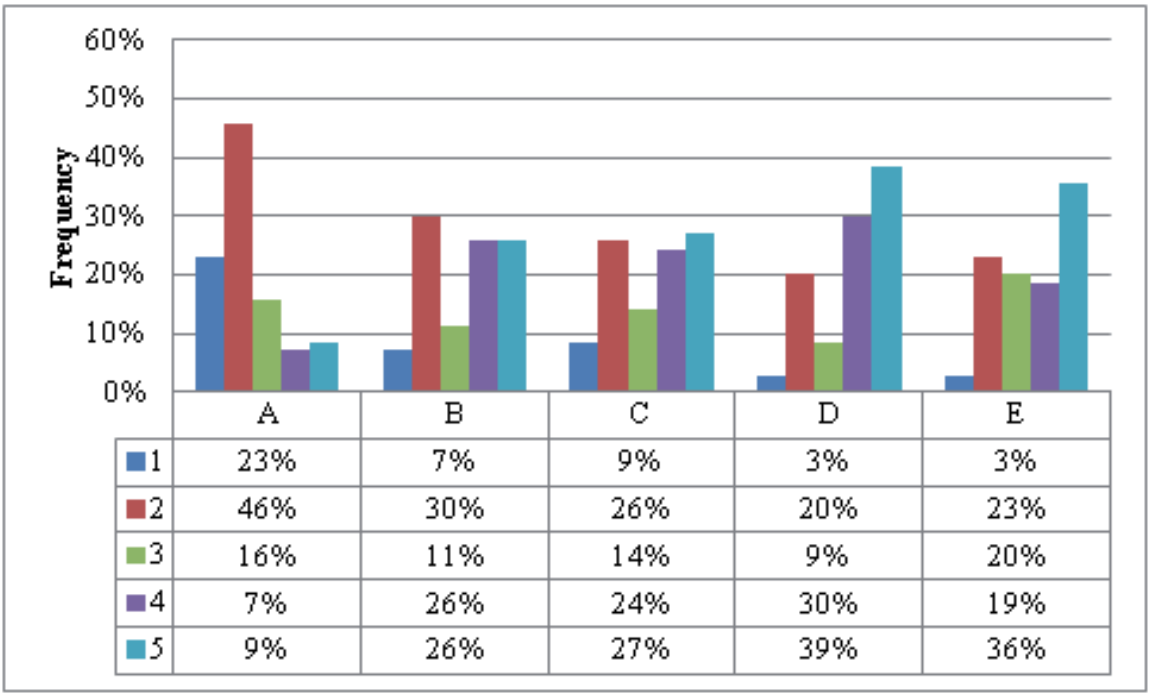
Figure 4.5. PE difficulty scores frequency

D obtained statistically better scores than B, C and A, even though the odds of getting higher scores in E was higher than in D. However, E did not present statistically significant differences with B and C. That is why D is considered to have the best scores in terms of PE difficulty (see Table A.3).

5.6. 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 4.6 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).

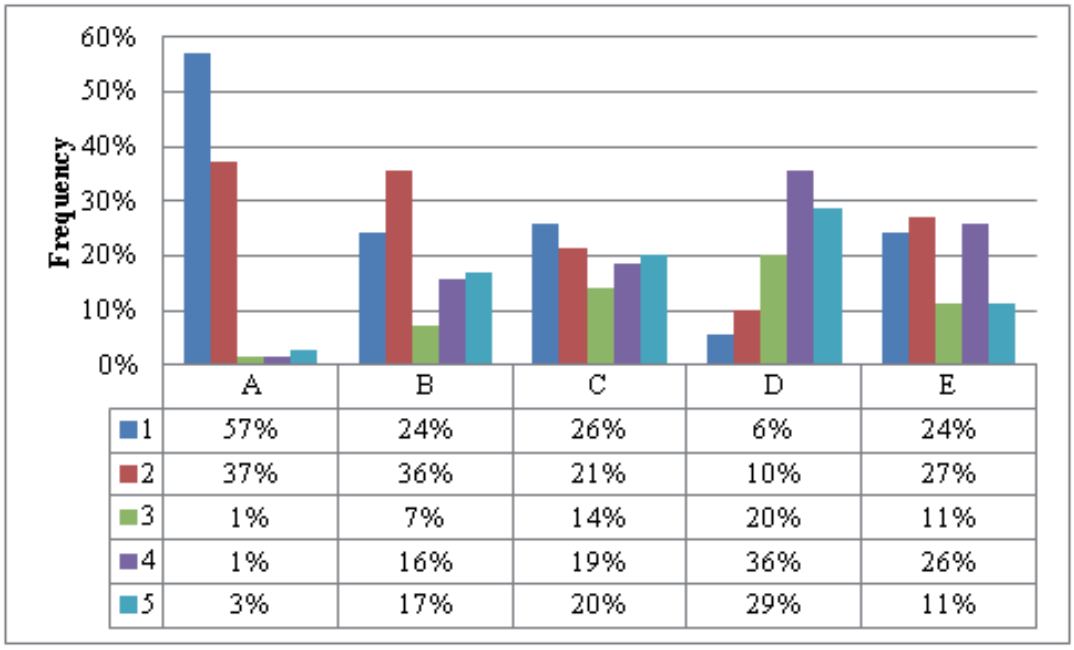


	A	B	C	D	E
1	23%	7%	9%	3%	3%
2	46%	30%	26%	20%	23%
3	16%	11%	14%	9%	20%
4	7%	26%	24%	30%	19%
5	9%	26%	27%	39%	36%

Figure 4.6. Adequacy scores frequency

5.7. Fluency

In terms of fluency, results were quite similar to those of adequacy. Figure 4.7 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).

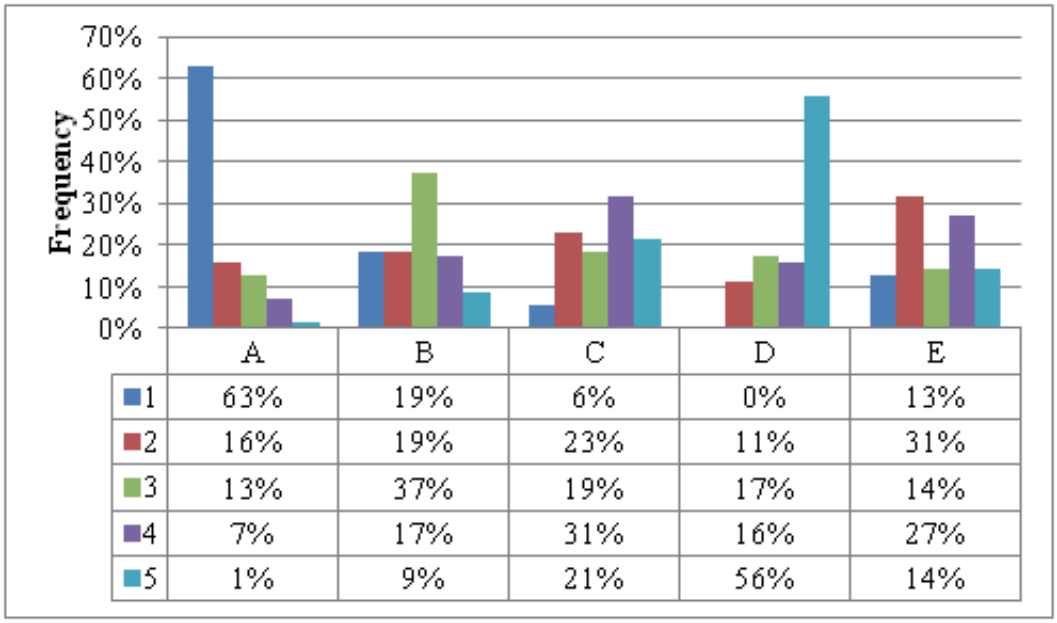


	A	B	C	D	E
1	57%	24%	26%	6%	24%
2	37%	36%	21%	10%	27%
3	1%	7%	14%	20%	11%
4	1%	16%	19%	36%	26%
5	3%	17%	20%	29%	11%

Figure 4.7. Fluency scores frequency

5.8. Ranking

According to Figure 4.8, 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 4.8. Ranking scores frequency

6. 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 3 of 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 in 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.

Reference List

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.

Appendixes

Table A.1. Selected clip for the test

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.2. PE necessity odds ratio (OR) table

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.3. PE difficulty 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.4. MT adequacy 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.5. MT fluency odds ratio (OR) table

Label	Pr > |t|	OR Estimate
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.6. Ranking odds ratio (OR) table

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

Chapter 5. Article 3: Machine Translation in audio description? Comparing creation, translation and post-editing efforts

5 Article 3

Fernández-Torné, A., Matamala, A. (2016). Machine Translation in Audio Description? Comparing Creation, Translation and Post-editing Efforts. *Skase*, *9*(1), 64-85.

Abstract

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

Keywords

Accessibility; audio description; audiovisual translation; machine translation; Catalan language; post-editing effort

1. Introduction

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

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

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

The article begins with an overview of related work. Next, the experimental set-up is presented, with a thorough description of the participants, test data, effort assessment methods, test development, and statistical methods used. In the following section, a

comprehensive exposition of the results is presented and discussed, and finally, conclusions are drawn while proposing directions for further research.

2. Related work

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

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

Measuring temporal effort is straightforward. Technical effort can also be directly observed by using methods such as key-logging technologies (Guerra 2003, Tatsumi and Roturier 2010). However, cognitive effort is not directly observable. Krings (2001) used think-aloud protocols to determine cognitive effort, but he noticed that this

method affected the total process time. Other technologies, such as key-logging (O'Brien 2004) and eye-tracking (O'Brien 2011, Carl et al. 2011), have successfully been used, since they allow subjects' behaviour to be recorded unobtrusively in real time. Pauses have been considered a key indicator of cognitive effort. Indeed, in writing research pauses are "assumed to provide us with a window to the cognitive processes underlying language production" (Wengelin 2006: 108, cited in Chukharev-Hudilainen 2014: 64), and they are usually computed, particularly their frequency, duration and position. Lacruz, Denkowski and Lavie (2014) state that both average pause ratio (APR), i.e. the average time per pause divided into the average time per word, and pause to word ratio (PWR), i.e. the number of pauses divided into the number of words, correlate well with cognitive effort: the lower the APR and the higher the PWR, the higher the levels of cognitive effort.

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

3. Methodology

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

3.1. Participants

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

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

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

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

3.2. Test data selection

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

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

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

3.3. Assessment measures

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

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

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

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

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

the more temporal effort involved in searching for information or solving doubts on the Internet.

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

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

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

3.4. Questionnaire design

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

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

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

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

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

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

a. In the AD creation PTQ:

- The clip was easy to audio describe.

b. In the AD translation PTQ:

- The source text was easy to translate.
- The clip was easy to audio describe departing from the original AD.

c. In the AD PE PTQ:

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

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

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

3.5. Test development

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

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

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

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

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

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

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

3.6. Statistical methods

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

4. Results and discussion

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

4.1. Objective effort

4.1.1. Temporal effort

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

The amount of time spent in Subtitle Workshop, where the actual task was to be performed, was also calculated. AD PE and AD translation presented a closer mean time (2,238.552 seconds, i.e. 37.31 minutes, and 2,245.303 seconds, i.e. 37.42 minutes, respectively) spent on the software, and for AD creation the time spent was

only slightly higher (2,415.218 seconds, 40.25). Again, the difference was not statistically significant.

When calculating relative values (see Figure 5.1), it was observed that in AD creation participants spent 90% of the time in Subtitle Workshop, which means it was the task requiring less research on the Internet, whereas AD translation was the task requiring most time outside Subtitle Workshop (33%). Post-editing was somewhere in between, dedicating 84% to the Subtitle Workshop and 16% to searching the Internet. These results can be seen as a logical consequence of the processes associated with each task: while AD can be considered a creative and introspective task, translation is usually associated with dictionary searches and online consultations. On the other hand, PE mainly implies rewording, word reordering and error correction, which do not necessarily involve as many Internet searches.

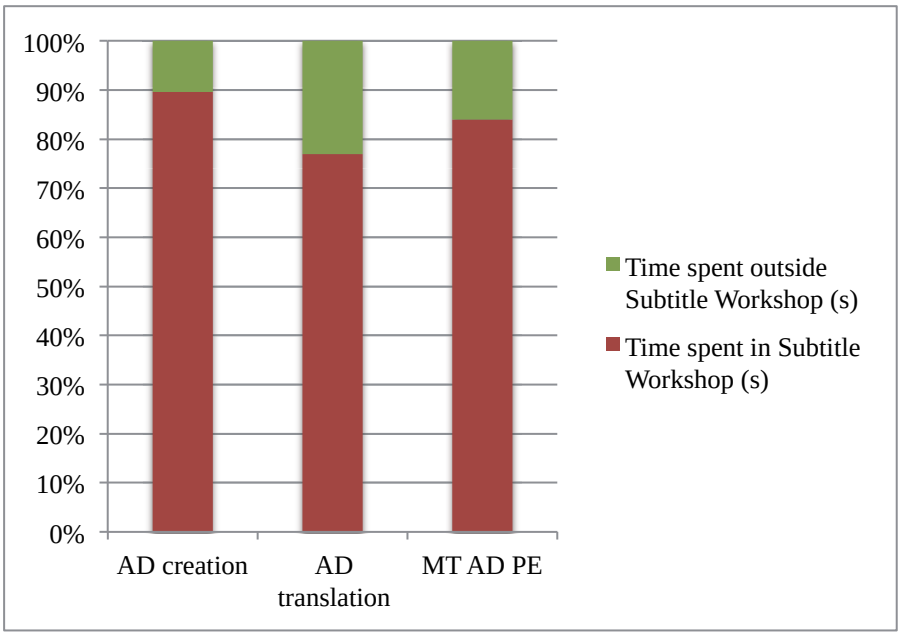


Figure 5.1 Time spent inside and outside Subtitle Workshop

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

4.1.2. Technical effort

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

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

Mouse actions presented quite similar mean figures: 1,473.667 for AD creation, 1,556.583 for AD PE and 1,666.545 for AD translation. In this respect, clicks and movements did not show significant differences either, but scrolls did (see Table 5.1). AD creation (23.417 scrolls on average) was statistically lower than both AD translation (65 scrolls) and post-editing (58.667 scrolls).

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

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

4.1.3. Cognitive effort

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

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

It was deemed interesting to see whether the distribution between the time spent pausing and the time devoted to active writing diverged from task to task. AD creation seemed to be assigning more time to pauses (60.27%), while AD translation and post-editing devoted just a little more than half of the time to pausing (51.53% and 52.29% respectively) (see Figure 5.2). Even though the difference was not significant, it is important to highlight that the task of creation involves more pausing than writing, which might be an indicator of a higher cognitive effort.

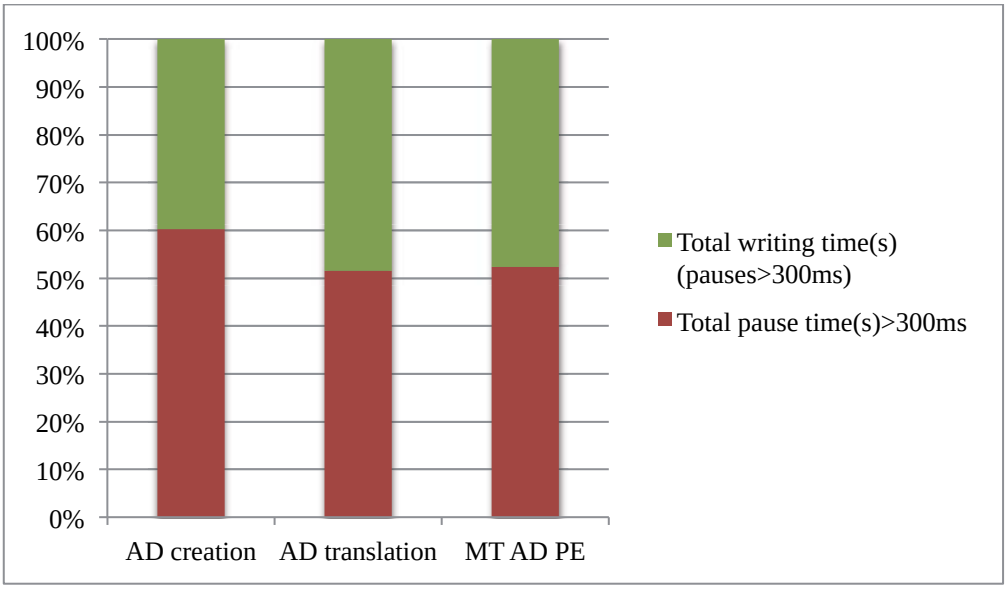


Figure 5.2 Distribution of pausing and writing during each task

All these data seem to indicate that post-editing was the least effort-involving task, especially if we focus on a key indicator such as PWR: AD PE presented the lowest number of pauses and the highest number of words, resulting in the lowest PWR, which is associated with low levels of cognitive effort. Conversely, AD creation seems to be the most demanding cognitively. Table 5.1 presents an overview of objective results.

		AD creation	AD translation	AD post-editing
Temporal effort	total process time (seconds)	2,696.880	2,919.641	2,666.695
		(44.95 minutes)	(48.66 minutes)	(44.44 minutes)
	time spent in Subtitle Workshop (seconds)	2,415.218	2,245.303	2,238.552
		(40.25 minutes)	(37.42 minutes)	(37.31 minutes)

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

Table 5.1 Overview of objective effort assessment results

4.2. Subjective effort and participants' opinions

Beyond objective effort, this research aimed to go a step further and gather data on participants' subjective views on effort and other relevant aspects. First of all, a comparison of the replies to the GQ, before and after the experiment, is presented,

focusing first on effort and then on other items such as the degree of creativity impairment each task involves, boredom, calque conveyance, and final quality. Secondly, the participants' opinions after each task are analysed, adopting a contrastive approach where possible.

4.2.1. General questionnaire responses

Quantitative data on the participants' expected effort (prior to the tasks) and perceived effort (after the task) was gathered through a questionnaire, which included also an open field to justify their choices in its post-task version. Opinions on other aspects were also gathered. Table 5.2 shows the means and medians obtained for each item under analysis before and after performing the tasks, on a 10-point scale where 1 is the lowest value. In the case of final quality, however, it must be clarified that "best quality" was number 1 whilst "worst quality" was number 10.

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

Table 5.2 Comparison of opinions before and after the experiment

Results indicate that participants expected AD PE to be the task that would impair their creativity the most and would convey more calques. They also expected it to be the most boring task, and the one delivering the worst output quality, and involving more effort. However, in terms of effort, once the experiment was finished, both AD creation and translation were perceived as involving less effort than expected (mean=8.25 and 6.17 prior to the test to 7.17 and 5.58 after the test), while PE AD showed the opposite trend (6.50 changed into 7.50), becoming the task involving more effort according to our sample of participants. Regarding the other indicators, they all showed a clear evolution towards worse PE ratings after performing the task. This was also the case for most indicators in AD creation and AD translation, except for creativity impairment in AD translation (7.45 prior to the test, 7.27 after), and boredom in both AD translation (4.18 both prior and after the test) and AD creation (2.09 into 1.82). One possible explanation for this trend is the lack of experience of our participants.

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

In connection with the degree of boredom of the tasks, responses reasserted that “[t]he AD creation task is the least boring task” (Participant 5) since “the more creative you can be, the less boring the activity will be” (Participant 7), which seems to indicate they enjoyed it more, although enjoyment was not directly assessed in the

questionnaire. They also agreed in terms of conveying calques that "[i]n both the translation and the MT AD post-editing you risk to use [*sic*] calques because you do not create a new text, but depart from a source text in a foreign language" (Participant 9), and that "MT AD lacks quality because the audiodescriber [*sic*] departs from a text which is not perfectly translated" (Participant 9).

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

4.2.2. Post-task questionnaires analysis

One of the questions included in all three post-task questionnaires assessed how easy participants felt a particular task was, immediately after performing it. Although not explicitly mentioning effort in the statement, this measure can somehow be linked to the effort participants perceived in the task. Figure 5.3 shows how many participants selected a specific value on a 5-point Likert scale for each task, 1 being "strongly disagree" and 5 being "strongly agree".

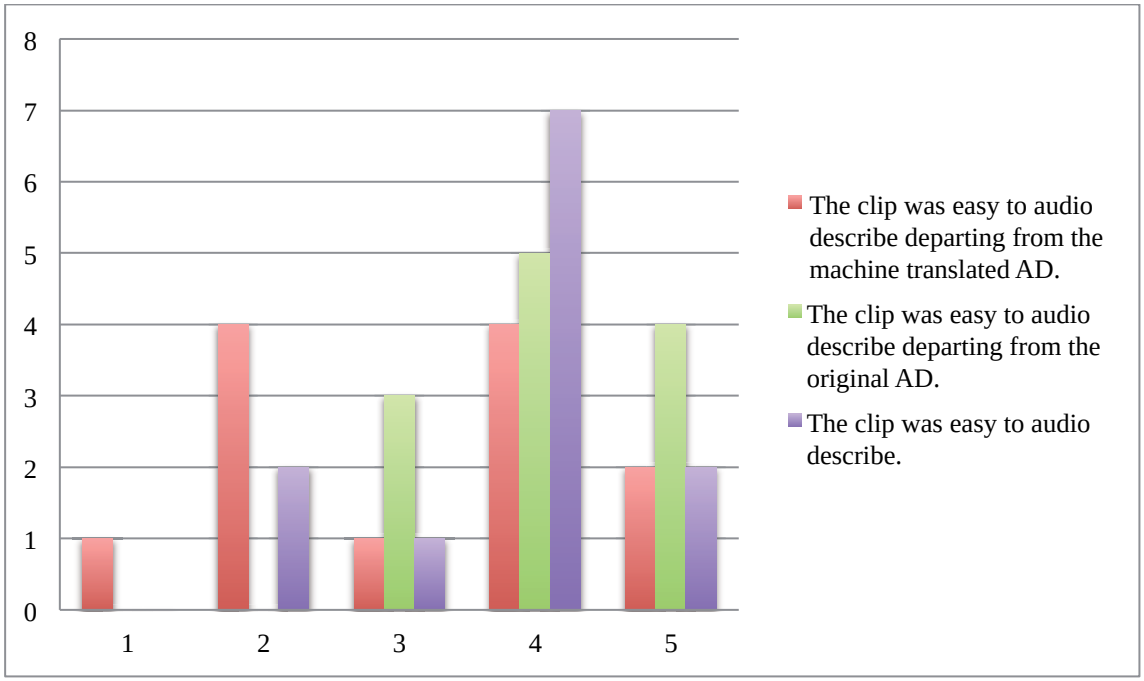


Figure 5.3. Self-reported ease of audio description in each scenario

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

5. Conclusions

This article has presented an experiment in which the efforts of creating an AD, translating an AD and post-editing a machine-translated AD were compared, with the

ultimate aim of exploring whether machine translation could be satisfactorily deployed in audio description. After presenting an overview of the current state of the art, and describing the experimental design, results were discussed.

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

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

Due to its exploratory nature, this experiment has several limitations, opening the door to further research. First of all, as already stated above – and despite it being common practice in this kind of research (Temizöz 2012) –, the small sample size does

not allow statistically robust conclusions to be drawn. A larger sample would allow for sounder extrapolations to be made.

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

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

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

Acknowledgements

We would like to thank Märielle Leijten and Eric Van Horenbeeck for their help and support with the keystroke-logging tool Inputlog. Special thanks are also due to the students of the MA in Audiovisual Translation who volunteered to take part in the experiments. This work was supported by the Spanish Ministerio de Ciencia e

Innovación project ALST (reference code FFI2012-31024). Anna Matamala is a TransMedia Catalonia member, funded by Generalitat de Catalunya (2014SGR27).

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

LACRUZ, Isabel, DENKOWSKI, Michael, LAVIE, Alon. 2014. Cognitive Demand and Cognitive Effort in Post-Editing [online]. In *Proceedings of the 3rd Workshop on Post-Editing Technology and Practice (WPTP3),* 2014 [cit. 2015-12-19], pp. 73-84. Available at: <http://amtaweb.org/AMTA2014Proceedings/AMTA2014Proceedings_PEWorkshop_fi

nal.pdf>.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 6. Summary

6 Summary

This PhD thesis explores the application of technologies to the audio description field with the aim to semi-automate the process in two ways. On the one hand, text-to-speech is implemented to the voicing of audio description in Catalan and, on the other hand, machine translation with post-editing is applied to the English audio descriptions to obtain Catalan AD scripts.

In relation to TTS, a selection of available synthetic and natural voices in Catalan (5 masculine ones and 5 feminine ones for each category) is assessed by means of a self-administered questionnaire mainly based on the ITU-T P.85 Standard Mean Opinion Score (MOS) scales for the subjective assessment of the quality of synthetic speech. Thus, participants assess the voices taking into account various items (overall impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort, and acceptance). The voices obtaining the best scores for each category are then used to assess the reception of text-to-speech audio descriptions compared to human-voiced audio descriptions by blind and visually impaired persons. Both quantitative and qualitative data obtained show that the preferential choice of blind and partially sighted persons is the audio description voiced by a human, rather than by a speech synthesis system since natural voices obtain statistically higher scores than synthetic voices. However, TTS AD is accepted by end users (94% of the participants) as an alternative acceptable solution, and 20% of the respondents actually state that their preferred voice from the four under analysis is a synthetic one.

As regards MT, a selection of five available free online machine translation engines from English into Catalan is evaluated in order to determine which is the most suitable for audio description. Their raw machine translation outputs and the post-editing effort involved are assessed using eight different scores, including human judgments (PE time, PE necessity, PE difficulty, MT output adequacy, MT output fluency and MT output ranking) and automatic metrics (HBLEU and HTER). The results show that there

are clear quality differences among the systems assessed and that one of them (Google Translate) is the best rated in six out of the eight evaluation measures used. Once the best performing engine is selected, the effort, both objective and subjective, involved in three scenarios is compared: the effort of creating an audio description from scratch (AD creation), of manually translating an audio description (AD translation), and of post-editing a machine-translated audio description (AD PE). The results show that the objective post-editing effort is lower than creating an AD *ex novo* and manually translating it, although the subjective effort is perceived to be higher for the post-editing task.

Chapter 7. Conclusions

7 Conclusions

As the demand for audio description is likely to grow in Europe due to the aging population and increasing legislation in media accessibility, alternative ways of creating and voicing AD need to be researched. It is within this framework that this thesis has investigated the semi-automatisation of the current AD process through the application of two mature technologies.

With this idea in mind, this thesis set two main objectives.

First, in relation to the recording of the AD, the implementation of TTS wanted to be explored by studying the reception of Catalan text-to-speech audio descriptions by blind and visually impaired persons, compared to traditional human-voiced audio descriptions. Taking into account that TTS is widespread in other fields such as screen readers, audiobooks, and GPS navigation devices, and that positive feedback was received from users in other languages, it was deemed that this research in Catalan was necessary.

In order to achieve this objective, a specific goal was set, i.e. to assess a selection of available synthetic and natural voices in Catalan, so as to determine which were the most adequate for the purposes of our research. To fulfil this aim, a methodological framework for the evaluation of voices needed to be developed.

Two hypotheses linked to this part of the research were proposed:

H1. Blind and visually impaired persons will accept text-to-speech audio descriptions.

H2. Blind and visually impaired users will prefer human voiced audio descriptions compared to text-to-speech audio descriptions.

Second, regarding the writing of the AD script, the implementation of MT with PE also wanted to be explored and so the inclusion of English-Catalan machine translation systems in the process of AD creation was researched. The processes of AD human creation and human translation were compared to the post-editing of MT AD,

focussing specifically on the effort involved in each process. This is especially relevant in a context in which the use of machine translation is expanding, as proved by the many EU-funded projects particularly in the subtitling domain.

In order to meet this main objective, a specific goal was established, namely to assess a selection of available free online machine translation engines from English into Catalan, so as to determine which one was the most adequate for our purposes in audio description.

Linked to the research on MT, two hypotheses were put forward:

H3. The effort of post-editing a machine translated audio description will be lower than that of creating an audio description from scratch.

H4. The effort of post-editing a machine translated audio description will be lower than that of translating it manually.

This chapter unfolds the most relevant findings of the studies performed and analyses whether the objectives are fulfilled while validating or refuting the hypotheses, differentiating between the studies around text-to-speech (section 7.1) and the research on machine translation as applied to audio description (section 7.2). Finally, it includes a section devoted to new pathways for future research (section 7.3).

7.1 Text-to-speech in audio description

For the evaluation of voices a quasi-experiment was designed following a quantitative research strategy and a within-subjects approach, as well as using questionnaires as the research method. Establishing the methodological framework for the evaluation protocol included various aspects. After deciding the number of voices to be assessed for each category (synthetic masculine, natural masculine, synthetic feminine and natural feminine), the particular voices to be used were selected from the available synthetic ones in Catalan and from volunteer voice talents in the case of the natural ones. It was decided that two different speech samples (one per genre) would be used

aking into account the experiment's design. The preparation of the speech samples for the specific assessment of voices in the AD application also implied determining the AV product from which to extract the AD units, the number of AD units per sample, the selection of the AD units, and the order of the AD units in each sample.

Regarding the elaboration of the questionnaires, it entailed selecting the items to be assessed and defining the order of presentation of the items, the question formulation, the answer labelling, the questionnaire delivery mode, and the data collection system. Based mainly on the ITU-T P.85 and P.800 recommendations, which develop methods for the subjective assessment of the quality of speech voice output devices, and later modified versions proposed by other authors (Viswanathan & Viswanathan, 2005; Hinterleitner et al., 2011; Vázquez & Huckvale, 2002), eight items were chosen to constitute the questionnaire used in the test, namely general impression, accentuation, pronunciation, speech pauses, intonation, naturalness, pleasantness, listening effort and acceptance. All these items were assessed on a 5-point scale, for which a mean opinion score (MOS) was calculated.

As for the participants, the sampling method, the sample size and the inclusion criteria for the sampling population to be recruited were established. Finally, the listening order of the speech samples was arranged.

The analysis of the quantitative data resulted in the selection of two synthetic and two natural voices (one best-rated for each category and genre): A (Laia, by Acapela) and H (Oriol, by Verbio) as artificial voices, and D (Belén Roca, a female professional voice talent) and H (Arià Paco, a male voice talent student) as natural voices. These four voices were then used to compare the reception of Catalan TTS and human-voiced audio descriptions by means of another within-subjects quasi-experiment involving blind and visually impaired patrons. Accordingly, end users assessed the quality of the four intervening AD voices as they perceived them using the same eight mean opinion score scales as in the pre-test.

The quantitative data gathered showed that natural voices obtained the highest mean scores. However, the results for the synthetic voices were very close to those of natural voices. Actually, all mean scores of the synthetic voices were above 3.1, reaching 4.313 in the accentuation of the synthetic male voice and 4.284 for the pronunciation of the feminine synthetic voice. All the median scores of both the natural and synthetic voices were between 3.0 and 5.0. Both natural voices obtained 4.0 in pleasantness and 5.0 in overall impression, accentuation, listening effort, naturalness, pronunciation and speech pauses, with the male natural voice getting higher scores than its feminine counterpart (5.0 vs 4.0) in acceptance and intonation. This shows how even a natural voice may not get the highest mark in terms of pleasantness or intonation, and how subjective aspects like acceptance present greater variation than standard features mastered by professional voice talents (e.g. accentuation and pronunciation).

The medians of the female synthetic voice were stable, obtaining 4.0 in all items under analysis, while the male artificial voice presented greater variability, although most items were rated 4.0 (acceptance, intonation, listening effort, speech pauses, pronunciation). Again, the fact that all items were assessed above 3.0 regardless of the nature of the voices and that the median scores were the same for natural and artificial voices in some items was particularly relevant.

Despite these positive evaluations, the statistical analysis also proved that there were significant differences between the synthetic voices and their natural counterparts in all items under analysis, the natural voices being considered better than the artificial ones in all cases.

The qualitative data gathered in the post-questionnaire supported this result. When asked whether they preferred to hear a human or a synthetic voice read the AD, 81% of the participants stated that they preferred a human voice, 1% declared that they preferred a synthetic voice, 3% said that it depended on the audiovisual product, and 15% declared they did not have any specific preferences as long as the artificial voice

sounded natural enough and was not tiring. It must be noted that in the case of the synthetic voices tested the naturalness mean scores were 3.507 for the female voice and 3.104 for the male one, and the listening effort mean scores were 3.836 and 3.657 respectively, which are quite strong results in a 5-point scale. It must also be stressed that 51 respondents (76%) stated that they use synthetic voice applications on a daily basis, which according to Cryer and Home (2009) may affect acceptability as they get used to them.

To complement the quantitative results obtained in the acceptance scales, participants were explicitly asked about the TTS AD being an alternative solution to human voiced audio description. 94% participants responded that they would accept it. Twenty-two of them, i.e. 33%, stated that the main reason for accepting it was that it would help increase the number of audio described audiovisual products, with eight of them further explaining that it would reduce both the costs and the time involved in their production. Nine participants (13%) stated that it could be an alternative solution because the quality of synthetic voices was already good enough, while 10 other informants (15%) stated that they would accept it simply because TTS AD was better than no AD at all, and nine respondents (13%) pointed out that it should only be an alternative, not the common situation. This is in line with the previous results obtained in Poland (Szarkowska, 2011).

Most participants agreed on applying TTS AD in documentaries (48 respondents), series (48 respondents) and films (49 respondents); 36 respondents would apply it to cartoons (36 respondents), and 24 would implement it in live plays. Only four participants were against implementing it at all. It remains to be seen, though, whether these subjective opinions would be translated into positive assessments in experiments involving a wide variety of genres.

Finally, a question about their opinion after listening to the four voices included in the experiment showed a preference for the masculine natural voice (42%) and the feminine natural voice (38%), although 14% said they prefer the feminine synthetic

voice and 6% selected the male synthetic voice. These qualitative data match with the results obtained both in the descriptive and inferential statistics, which actually graded voices in the same order: the natural masculine voice was the one which obtained better mean scores, closely followed by the natural feminine, then the synthetic feminine and finally the synthetic masculine.

Thus, the results obtained in the reception study led to the validation of the hypotheses 1 and 2 in relation to TTS. On the one hand, results showed that blind and visually impaired persons accepted text-to-speech audio descriptions, while on the other they also showed that they preferred human voiced audio descriptions.

All in all, this thesis has provided results on the evaluation of text-to-speech audio description in Catalan and, inspired by previous works and existing methodological frameworks, it has also developed an evaluation methodology for the assessment of both synthetic and natural voices in audio description which has been implemented in a selection of voices. Hence, the secondary aim of assessing a selection of voices was met, allowing for the fulfilment of the main objective and the validation of the first two hypotheses.

7.2 Machine translation in audio description

The second main objective of this thesis was to research the inclusion of English-Catalan machine translation systems in the process of AD creation, compared to the process of human creation and human translation, focussing specifically on the effort involved in each process. Again, in order to attain this objective, the specific goal of assessing a selection of available free online machine translation engines from English into Catalan was set to determine which was the most adequate for our purposes in audio description. In order to obtain a best-rated machine translation engine, a methodological framework also needed to be established.

Thus, a methodology for the quality evaluation of different MT engines in the AD domain was developed based on both objective and subjective measures. Eight

different scores were established to assess the quality of their raw machine translation outputs and the effort which their post-editing involved, namely HBLEU and HTER as automatic metrics, and PE time, PE necessity, PE difficulty, MT accuracy, MT fluency and MT ranking, as human measures. These measures were selected after a thorough literature review (Callison-Burch et al., 2012; Federmann, 2012; Popovic et al., 2013; Specia, 2011; De Sousa et al., 2011; Chatzitheodorou & Chatzistamatis, 2013; Koehn & Monz, 2006; Koponen, 2010).

HBLEU measured how close was the MT to its post-edited versions (Del Pozo, 2014) and HTER measured the distance 'between machine translations and their post-edited versions' (Specia, 2011, p. 74). While PE time referred to the total time spent in the post-editing of each AD unit, 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 and PE difficulty referred to how difficult post-editing the raw MT output had been. MT adequacy assessed whether all the content of the source text was included in the target text, MT fluency judged to what extent a translation flowed naturally and was considered genuine in the target language, and the ranking classified AD units according to their global quality.

Thus, the best MT system should be the one obtaining the highest HBLEU, the lowest HTER, the lowest PE time, and the highest position in the ranking. In relation to the remaining subjective human assessments, they were presented 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. Thus, the best MT system should also present the highest PE necessity, PE difficulty, MT adequacy and MT fluency scores.

The engines under study were Yandex Translate, by Yandex; Google Translate, by Google; Apertium, by Universitat d'Alacant; Lucy Kwik Translator, by Lucy Software and

Services GmbH; and Bing Translator, by Microsoft.

The analysis of the automatic and human measurements resulted in the determination of one best-rated MT engine. Google Translate was found to be the best 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 adequacy presenting higher scores than three of the remaining MT systems. In relation to the objective assessments, Google Translate also obtained the highest HBLEU scores and outperformed three of 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, probably due to the limited length of the source text (240 words).

Google Translate was then used to compare the effort, both objective and subjective, involved in three scenarios: creating an audio description from scratch (AD creation), manually translating an audio description (AD translation), and post-editing a machine-translated audio description (AD PE).

For the assessment of the effort, Krings's approach (2001) was followed and thus temporal effort, technical effort and cognitive effort were distinguished. Several parameters from each category were measured, namely total process time, time spent in the software used (Subtitle Workshop) for the temporal effort; keyboard actions (including total character types and other keystrokes), mouse actions (including clicks, movements and scrolls), switches keyboard to mouse, and total window transitions, for the technical effort; and total pause time, mean pause time, number of pauses and pause-to-word ratio, for the cognitive effort. Data were also gathered via questionnaires to assess the participants' subjective effort.

The results of such a comparison brought about the validation of hypotheses 3 and 4. Thus, it was proved that the effort of post-editing a machine translated audio description is lower than that of creating an audio description from scratch and that of translating it manually. However, it must be noted that no statistically significant

differences were found for many measures, probably due to sample size limitations. In fact, wider differences in terms of the efforts involved by each task were expected, particularly in the time that each task demanded, in line with the findings reported in De Sousa, Aziz, and Specia (2011) in relation to subtitles. In our test, despite PE being the fastest option, the differences were extremely low: on average post-editing took only four minutes less than translating, and the difference between post-editing and creating an AD was just a few seconds. In relation to technical effort, post-editing was statistically the least keyboard intensive task, with significantly the lowest number of characters typed, in accordance with O'Brien's (2010) findings. It was also the task entailing fewer mouse clicks and fewer switches from keyboard to mouse, while the rest of the values were not the highest for the three tasks in any case. In terms of cognitive effort, post-editing presented the lowest pause-to-word ratio, which is associated with low levels of cognitive effort. Conversely, AD creation was the most cognitively demanding, showing statistically significant differences with post-editing in this sense.

However, the data obtained in the questionnaires indicated that the participants' subjective effort was higher for the PE task. Actually, and contrary to what we expected, the subjective perceived effort, i.e. after the PE task (mean 7.50; median 8), was even higher than the subjective expected effort, i.e. before performing the PE task (mean 6.50; median 6). These facts might be due to the participants' reluctance to the use of MT in creative environments such as AD and to their lack of experience in MT post-editing. Indeed, participants not only expected AD PE to be the task involving more effort, but also the one that would impair their creativity the most, that would convey more calques, that would be most boring, and that would deliver the worst output quality, which provide evidence of their reluctance before the task. Such reluctance increased after performing the task as all indicators showed a clear evolution towards worse perceived ratings. Although this tendency to worse ratings after performing the tasks was also shown in most indicators in AD creation and AD translation, both AD creation and translation were perceived as involving less effort

than expected (from 8.25 and 6.17 prior to the test to 7.17 and 5.58 after the test, respectively), AD creation was perceived as being less boring (from 2.09 to 1.82), and AD translation as being less creativity impairing than expected (from 7.45 to 7.27).

Again, this thesis has methodologically contributed to the MT evaluation sphere. Although it focused on the audio description domain in a particular language combination (from English to Catalan), this experimental procedure may be replicated to assess the quality of MT engines in any given domain and linguistic pair.

Another major contribution is the effort assessment in audio description. Rather than putting the emphasis on the development of any specifically trained MT, it has focused on the describer's/translator's relationship with this technology. If the use of MT technology is to be introduced in audio description, It is our belief that extensive user testing like the one done in this PhD should be carried out to prove that it is worth it by comparing the effort involved in writing AD scripts from scratch, manually translating AD scripts from another language and post-editing the MT AD scripts. Concepts initially developed in relation to post-editing (temporal, technical and cognitive efforts) have been applied to assess the effort of text production and translation in this audiovisual transfer mode, allowing for an objective comparison of the efforts involved in each scenario. Although with a longer tradition in translation process research applied to other translation domains, the use of tools such as keyboard logging is also an innovation to a certain point in audiovisual translation and media accessibility studies.

7.3 Future research

There are many new research pathways that aroused while carrying out the experiments. For the sake of narrowing down the scope of the research, several decisions needed to be taken leaving aside many other possibilities which would definitely be worth investigating. On the other hand, some other avenues for future research emerge from this thesis' limitations due to many factors.

First, in the TTS pre-test both sighted and blind and visually impaired participants could be used to ascertain whether the perception of both groups of participants is comparable in terms of voice preferences. It would also be worth researching whether their previous experience with TTS voices has an impact on their choices.

As far as the TTS experiment is concerned, the clips chosen belonged to one film representing a miscellaneous genre (Salway et al., 2004). Selecting clips belonging to other various genres and examining whether results would vary depending on the genre would confer higher reliability on the conclusion that TTS is accepted by the blind and partially sighted. Moreover, the experimental design required that the length of the clips was limited to around three minutes. It would be interesting to test whether the results, particularly in terms of acceptance, would remain the same in longer productions as fatigue could play a role in the assessment of voices. As a researcher, a decision had to be made on whether to prioritize an experiment in a controlled environment or in a real-world setting such as a cinema, where the playing of longer productions would have been a better option, and a controlled experiment was preferred as a first step.

It would also be highly interesting to see whether reception varies in productions originally shot in Catalan and in dubbed productions, since the language and the sound conditions are different. In this regard, it may well be that the sound of non-dubbed audiovisual products negatively affects the acceptance of TTS audio description as would find it more shocking.

In line with other research studies carried out by different authors, it would also be worth researching whether the fact that the AD is voiced by a speech synthesiser influences the engagement of the audience (Fryer & Freeman, 2013b), whether it affects the end users comprehension (Cabeza-Cáceres, 2013), and whether it has an impact on their ability to recall the AD content (Fresno, 2014). End users reception could also be studied if they were given the possibility of tuning their own AD preferences. The tuning could involve selecting the voice gender, the voice and the

volume, in keeping with Walczak and Szarkowska's approach (2012), or the AD style, be it standard vs. extended audio descriptions, following Kobayashi et al.'s work (2010), or standard vs AD drafted according to sighted viewers' perception, in line with Mazur and Chmiel (2016).

In relation to the MT pre-test, although this was approached as a test previous to the main experiment in which a small sample of five participants was preferred to a subjective decision by the researcher, increasing the number of participants would be desirable to obtain a sounder statistical analysis of the results. In this sense, the professional profile of the participants could also be taken into account to see whether 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.

Expanding the AD data sources, such as other film genres, series and documentaries, would also improve the reliability of the test results. Moreover, apart from HBLEU and HTER, the MT evaluation model proposed could include other automatic metrics for the sake of balancing automatic metrics and human assessments, and an analysis of the correlation between the various measures would undoubtedly yield interesting results in the field of machine translation evaluation.

Synchronising (time-coding) and adjusting the post-edited AD units should also be considered in the assessment of the PE effort in AD, since it is an essential part of it. In this sense, a PE tool with audiovisual capabilities would be needed for the professional practice of AD PE. This tool should need to include the original time-coded AD and the raw MT AD. It should have a text editor in which to post-edit the MT version and a player in which images should be seen, together with the AD units in the form of subtitles for the audio describer to adjust them to silent gaps, similar to current subtitling tools.

Another interesting research avenue would be to test the performance of Google Translate in AD, the best-rated MT engine according to our evaluation, compared to other MT systems specifically trained with data belonging to the AD domain, as

training the engines has been proven to have better results in terms of quality (Armstrong et al., 2006; Volk, 2008). To do that, however, an English-to-Catalan AD corpus should be compiled, which nowadays is non-existent as the practice of AD is creating ADs from scratch. In the absence of an AD translation corpus, the translations of the audiovisual products' scripts could be used to feed the MT systems.

Finally, in reference to the MT experiment, again a larger number of participants would allow for sounder extrapolations to be made from the results. Taking into consideration the participants' profile, which was decided to be controlled to cater for a homogeneous sample in terms of AD experience and educational background, would also be interesting to see whether more experience in translation, post-editing or audio description has an impact on the results. It would also be interesting to find out whether there would be any differences between professionals with different profiles (audio describers, translators, post-editors). Additionally, it would be important to see whether professionals would spend less time performing the tasks than novices, in line with Moorkens and O'Brien's findings (2015). Comparing the performance of professionals with novices could also provide some interesting results for trainers in the field.

Apart from evaluating the process, a necessary step in future research would be the evaluation of the output quality obtained in each case. Studying the reception of AD quality and AD acceptance both by AD experts and by blind and visually impaired audiences would bring about very interesting results. They may conclude that, even though the AD post-editing task was the shortest, the resulting AD was the one obtaining the lowest quality and acceptance scores. This is in line with the research by Ortiz-Boix and Matamala (2015). Also, it would be worth testing whether the presumable difference in styles and lengths of the resulting ADs would affect the end users assessment of the AD quality and acceptance.

Finally, it would be worthwhile replicating the same experiment with other data sets and language pairs, to get a wider overview of the possibilities of machine translation

in this new field. Many research possibilities emerge, but this thesis can be considered a first step in the study of the semi-automatisation of the AD production process.

Updated bibliography

Updated bibliography

The reference lists and bibliography sections in the articles follow different styles, according to the guidelines established by each journal in this respect. In order to provide this thesis with a unified format, the APA Formatting and Style Guide has been used for all references in the thesis. Additionally, the updated references of the works cited in the articles will be offered, as some time has passed since their publication.

Alías, F., Iriondo, I., & Socoró, J. (2011). Aplicació de tècniques de generació automàtica de la parla en producció audiovisual. *Quaderns Del CAC*, *37*(1), 105-114. Retrieved from http://www.cac.cat/pfw_files/cma/recerca/quaderns_cac/Q37_Alias_etal.pdf

Arma, S. (2011). *The Language of Filmic Audio Description: A Corpus-Based Analysis of Adjectives* (Unpublished MA thesis). Università degli Studi di Napoli Federico II, Italy. http://dx.doi.org/10.6092/UNINA/FEDOA/8740

Armstrong, S., Caffrey, C., & Flanagan, M. (2006). Translating DVD subtitles from English-German and English-Japanese using example-based machine translation. In *MuTra - Multidimensional Translation Conference Proceedings. Audiovisual Translation Scenarios* (no page numbers). Copenhagen, Denmark: MuTra. Retrieved from http://www.euroconferences.info/proceedings/2006_Proceedings/2006_proceedings.html

Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D., & O'Hagan, M. (2006). Improving the quality of automated DVD subtitles via example-based machine translation. In *Translating and the Computer 28: Proceedings of the Twenty-Eighth International Conference on Translating and the Computer* (no page numbers). London, UK: ASLIB. Retrieved from http://www.mt-archive.info/Aslib-2006-Armstrong.pdf

Asakawa, C., & Takagi, H. (2007). Text Entry for People with Visual Impairments. In S. MacKenzie, I. & K. Tanaka-Ishii, *Text Entry Systems* (pp. 305-318). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Asociación Española de Normalización y Certificación. (2005). *Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías. UNE 153020*. Madrid, Spain: AENOR.

Aziz, W., de Sousa, S., & Specia, L. (2012). PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3982-3987). Istanbul: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/index.html

Bardini, F. (2015). *Audio description and film experience: Design of a reception study*. Poster, Fifth Advanced Research Seminar on Audio Description (ARSAD), Barcelona, Spain. Retrieved from http://grupsderecerca.uab.cat/arsad/sites/grupsderecerca.uab.cat.arsad/files/bardini_ARSAD_2015.pdf

Bowker, L. & Fisher, D. (2010). Computer-aided translation. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of Translation Studies* (pp. 60-65). Amsterdam, The Netherlands: John Benjamins.

Braun, S. (2008). Audio description research: state of the art and beyond. *Translation Studies In The New Millennium*, *6*, 14-30. Retrieved from http://epubs.surrey.ac.uk/303022/1/fulltext.pdf

Braun, S. (2011). Creating Coherence in Audio Description. *Meta*, *56*(3), 645-662. http://dx.doi.org/10.7202/1008338ar

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford, UK: Oxford University Press.

Cabeza-Cáceres, C. (2013). *Audiodescripció i recepció. Efecte de la velocitat de narració, l'entonació i l'explicitació en la comprensió fílmica* (PhD Thesis). Universitat Autònoma de Barcelona, Spain.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation* (pp. 10-51). Montréal, Canada: Association for Computational Linguistics. Retrieved from http://www.statmt.org/wmt12/WMT-2012.pdf

Carl, M., Dragsted, B., Elimng, J., Hardt, D., & Jakobsen, A. (2011). The Process of Post-Editing: a Pilot Study. In *Proceedings of the 8th International NLPSC Workshop* (pp. 131-142). Copenhagen, Denmark: Copenhagen Studies in Language. Retrieved from http://bridge.cbs.dk/events/CSL_41_complete.pdf

Carl, M., Jakobsen, A., & Jensen, K. (2008). Modelling human translator behaviour with user-activity data. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation (EAMT)* (pp. 21-26). Hamburg, Germany: Hitec.

Caruso, B. (2012). Audio Description Using Speech Synthesis. In *9th International Conference on Language Transfer in Audiovisual Media. 9th International Conference on Language Transfer in Audiovisual Media* (pp. 59-60). Berlin, Germany: ICWE.

Chapdelaine, C. & Gagnon, L. (2009). Accessible Videodescription On-Demand. In New York City, NY, New York City, NY, *'09. Proceedings of the 11th international ACM SIGACCESS Conference on Computers and Accessibility* (pp. 221-222). New York City, NY, USA: ACM. http://dx.doi.org/10.1145/1639642.1639685

Chatzitheodorou, K. (2013). COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. *The Prague Bulletin Of Mathematical Linguistics*, *100*(1), 83-89. http://dx.doi.org/10.2478/pralin-2013-0014

Chmiel, A. & Mazur, I. (2012). AD reception research: Some methodological considerations. In E. Perego (Ed.), *Emerging topics in translation: Audio description* (1st ed., pp. 57-80). Trieste, Italy: EUT Edizioni Università di Trieste.

Choudhury, R. & McConnell, B. (2013). *Translation technology landscape report*. De Rijp, The Netherlands: TAUS BV. Retrieved from https://www.taus.net/think-tank/reports/translate-reports/taus-translation-technology-landscape-report

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal Of Writing Research*, *6*(1), 61-84. http://dx.doi.org/10.17239/jowr-2014.06.01.3

Corporació Catalana de Mitjans Audiovisuals. (2013). *TV3 millora l'accés als continguts de les notícies als invidents*. Retrieved from http://www.ccma.cat/324/TV3-millora-lacces-als-continguts-de-les-noticies-als-invidents/noticia/2174075/

Cryer, H. & Home, S. (2008). *Exploring the use of synthetic speech by blind and partially sighted people*. Birmingham, UK: RNIB Centre for Accessible Information.

Cryer, H. & Home, S. (2009). *User attitudes towards synthetic speech for Talking Books*. Birmingham, UK: RNIB Centre for Accessible Information.

Cryer, H., Home, S., & Morley Wilkins, S. (2010). *Synthetic speech evaluation protocol*. Birmingham, UK: RNIB Centre for Accessible Information.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2015). The impact of machine translation error types on post-editing effort indicators. In *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)* (pp. 31-45). Miami, FL, USA: AMTA. Retrieved from http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf

Dam-Jensen, H. & Heine, C. (2013). Writing and Translation process research: Bridging the gap. *Journal Of Writing Research*, *5*(1), 89-101. http://dx.doi.org/10.17239/jowr-2013.05.01.4

De Jong, F. (2006). Access Services for Digital Television. In R. Pérez-Amat & Á. Pérez-Ugena (Eds.), *Sociedad, integración y televisión en España* (pp. 331-344). Madrid, Spain: Laberinto Comunicación.

De Sousa, S., Aziz, W., & Specia, L. (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 2011* (pp. 97-103). Hissar, Bulgary: RANLP. Retrieved from http://aclweb.org/anthology/R11-1014

Del Pozo, A. (2014). *SUMAT final report*. Donostia, Spain: Vicomtech.

Delgado, H., Matamala, A., & Serrano, J. (2015). Speaker diarization and speech recognition in the semi-automatization of audio description: An exploratory study on future possibilities?. *Cad. De Trad.*, *35*(2), 308-324. http://dx.doi.org/10.5007/2175-7968.2015v35n2p308

Denkowski, M. & Lavie, A. (2012). TransCenter: Web-based translation research suite. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)* (no page numbers). San Diego, CA, USA: AMTA. Retrieved from http://amta2012.amtaweb.org/AMTA2012Files/html/

Derbring, S., Ljunglöf, P., & Olsson, M. (2009). SubTTS: Light-weight automatic reading of subtitles. In *Nodalida'09: Proceedings of the 17th Nordic Conference of Computational Linguistics* (pp. 272-274). Odense, Denmark: Northern European Association for Language Technology. Retrieved from http://dspace.ut.ee/bitstream/handle/10062/9805/paper77.pdf?sequence=1&isAllowed=y

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT '02: Proceedings of the Second International Conference on Human Language Technology Research* (pp. 138-145). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Drożdż-Kubik, J. (2011). *Harry Potter i Kamień Filozoficzny słowem malowany – czyli badanie odbioru filmu z audiodeskrypcją z syntezą mowy* (Unpublished MA thesis). Jagiellonian University, Poland.

Encelle, B., Ollagnier-Beldame, M., Pouchot, S., & Prié, Y. (2011). Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis. In *ASSETS '11: Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 123-130). New York City, NY, USA: ACM.

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., & van Loenhout, G. et al. (2014). Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC 2014* (pp. 46-53). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from http://file:///Users/AFT/Downloads/LREC2014_Proceedings/pdf/463_Paper.pdf

European Commission. *MT@EC machine translation solution - ISA - European Commission*. *Ec.europa.eu*. Retrieved 23 May 2016, from http://ec.europa.eu/isa/ready-to-use-solutions/mt-ec_en.htm

European Union Agency for Fundamental Rights (FRA),. (2014). *Accessibility standards for audio-visual media*. Vienna: FRA. Retrieved from http://fra.europa.eu/en/publications-and-resources/data-and-maps/comparative-data/political-participation/audiovisual-standards

Federmann, C. (2012). Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin Of Mathematical Linguistics*, *98*(1), 25-35. http://dx.doi.org/10.2478/v10108-012-0006-9

Fernández-Torné, A. (Forthcoming). Machine Translation Evaluation through Post-Editing Measures in Audio Description. *inTRAlinea*, *18*.

Fernández-Torné, A. & Matamala, A. (2013). *Methodological considerations for the evaluation of TTS AD's acceptance in the Catalan context*. Presentation, Fourth Advanced Research Seminar on Audio Description (ARSAD), Barcelona, Spain. Retrieved from http://ddd.uab.cat/record/117078

Fernández-Torné, A. & Matamala, A. (2015). Text-to-Speech vs Human Voiced Audio Descriptions: A Reception Study in Films Dubbed into Catalan. *The Journal Of Specialised Translation*, *24*, 61-88. Retrieved from http://www.jostrans.org/issue24/art_fernandez.pdf

Fernández-Torné, A. & Matamala, A. (2016). Machine Translation in Audio Description? Comparing Creation, Translation and Post-editing Efforts. *Skase*, *9*(1), 64-85.

Fernández-Torné, A., Matamala, A., & Ortiz-Boix, C. (2012). *Technology for accessibility in multilingual settings: the way forward in AD?* Presentation, The translation and reception of multilingual films, Montpellier, France. Retrieved from http://ddd.uab.cat/record/117160

Freitas, D. & Kouroupetroglou, G. (2008). Speech technologies for blind and low vision persons. *Technology and Disability, 20*, 135-156.

Fresno, N. (2014). *La (re)construcción de los personajes fílmicos en la audiodescripción. Efectos de la cantidad de información y de su segmentación en el recuerdo de los receptores* (PhD Thesis). Universitat Autònoma de Barcelona, Spain.

Fryer, L. & Freeman, J. (2012). Presence of those with and without sight: Audio description and its potential for virtual reality applications. *Journal of CyberTherapy & Rehabilitation, 5*(1), 15-23.

Fryer, L. & Freeman, J. (2013a). Cinematic language and the description of film: keeping AD users in the frame. *Perspectives*, *21*(3), 412-426. http://dx.doi.org/10.1080/0907676x.2012.693108

Fryer, L. & Freeman, J. (2013b). Visual impairment and presence: measuring the effect of audio description. In *Proceedings of the 2013 Inputs-Outputs Conference:*

An Interdisciplinary Conference on Engagement in HCI and Performance (no page numbers). New York City, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/2557595.2557599

Garcia, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation*, *25*(3), 217-237. http://dx.doi.org/10.1007/s10590-011-9115-8

Garcia, L. (2004). Assessment of text reading comprehension by Spanish-speaking blind persons. *British Journal Of Visual Impairment*, *22*(1), 4-12. http://dx.doi.org/10.1177/026461960402200102

Georgakopoulou, Y. (2010). Challenges for the audiovisual industry in the digital age: Accessibility and multilingualism. In *Proceedings of META Forum 2010* (no page numbers). Brussels, Belgium: META-NET.

Georgakopoulou, Y. (2011). Challenges for the audiovisual industry in the digital age: The ever-changing needs of subtitle production. *The Journal Of Specialised Translation*, *17*, 78-103. Retrieved from http://www.jostrans.org/issue17/art_georgakopoulou.pdf

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2016). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 33-41). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from https://aclweb.org/anthology/W/W13/W13-2300.pdf

Groves, D. (2011). *MT at the CNGL*. Presentation, TAUS User Conference 2011, Santa Clara, CA, USA. Retrieved from https://www.youtube.com/watch?v=QJcZ97GEmDg

Guerberof, A. (2009). Productivity and quality in MT post-editing. In *MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT* (no page numbers). Ottawa, Ontario, Canada: AMTA. Retrieved from http://www.mt-archive.info/MTS-2009-Guerberof.pdf

Guerra, L. (2013). *Human translation versus machine translation and full post-editing of raw machine translation output* (Unpublished MA Thesis). Dublin City University, Ireland.

Hailes, S. (2013). *Improving Accessibility: The Emergence Of Spoken Subtitles*. Retrieved from

http://www.4rfv.co.uk/industrynews/170020/improving_accessibility_the_emergence_of_spoken_subtitles

Hanoulle, S., Hoste, V., & Remael, A. (2015). The efficacy of terminology-extraction systems for the translation of documentaries. *Perspectives*, *23*(3), 359-374. http://dx.doi.org/10.1080/0907676x.2015.1010549

Hinterleitner, F., Neitzel, G., Möller, S., & 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* (no page numbers). Turin, Italy: International Speech Communication Association (ISCA). Retrieved from http://festvox.org/blizzard/bc2011/DeutcheTelekom_Blizzard2011.pdf

Housley, J. (2012). *Ruqual: A system for assessing post-editing* (PhD Thesis). Brigham Young University, USA. Retrieved from http://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=4105andcontext=etd

Huang, X., Acero, A., & Hon, H. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. New Jersey, USA: Prentice Hall PTR.

Hyks, V. (2005). Audio description and translation: Two related but different skills", *Translating Today Magazine*, *4*(1), 6–8.

International Telecommunication Union. Telecommunication Standardization Sector (1994). *Telephone transmission quality subjective opinion tests. A method for subjective performance assessment of the quality of speech voice output devices. ITU-T Recommendation P.85. Geneva*, Switzerland: ITU. Retrieved from http: //www.itu.int/rec/T-REC-P.85-199406-I/en

International Telecommunication Union. Telecommunication Standardization Sector (1996). *Telephone transmission quality. Methods for subjective determination of transmission quality. ITU-T Recommendation P.800***.** Geneva, Switzerland: ITU. Retrieved from http: //www.itu.int/rec/T-REC-P.85-199406-I/en

Jacobs, S. (2011). *Apps4Android's Text-to-Speech based Audio Description Application for You Tube Videos*. Retrieved from https://www.youtube.com/watch?v=yiQq8KeMIIA

Jankowska, A. (2013). *Tłumaczenie skryptów audiodeskrypcji z języka angielskiego jako alternatywna metoda tworzenia skryptów audiodeskrypcji [Translation of audio description scripts from English as an alternative method of audio description scripts creation]* (PhD Thesis). Jagiellonian University, Poland.

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

Jekat, S. (2015). *Evaluation of audio description*. Presentation, Fifth Advanced Research Seminar on Audio Description (ARSAD), Barcelona, Spain. Retrieved from http://grupsderecerca.uab.cat/arsad/sites/grupsderecerca.uab.cat.arsad/files/Jekat_ARSAD_2015.pdf

Kobayashi, M., Fukuda, K., Takagi, H., & Asakawa, C. (2009). Providing synthesized audio description for online videos. In *ASSETS '09. Proceedings of the 11th international ACM SIGACCESS Conference on Computers and Accessibility* (pp. 249-250). New York City, NY, USA: ACM. http://dx.doi.org/10.1145/1639642.1639699

Kobayashi, M., O'Connell, T., Gould, B., Takagi, H., & Asakawa, C. (2010). Are Synthesized Video Descriptions Acceptable? In *ASSETS '10: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 163-170). New York City, NY, USA: ACM.

Koehn, P. & Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation* (pp. 102-121). New York City, NY, USA: Association for Computational Linguistics.

Koglin, A. (2015). An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors. *Translation and Interpreting*, *7*(1), 126-141. Retrieved from http://www.trans-int.org/index.php/transint/issue/view/29

Koponen, M. (2010). Assessing machine translation quality with error analysis. In *MikaEL: Electronic proceedings of the KäTu symposium on translation and interpreting studies 4*, Retrieved from https://sktl-fi.directo.fi/@Bin/40701/Koponen_MikaEL2010.pdf

Koponen, M. (2015). How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)* (pp. 2-15). Miami, FL, USA: AMTA. Retrieved from http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf

Koponen, M., Ramos, L., Aziz, W., & Specia, L. (2012). Post-Editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)* (pp. 11-20). San Diego, CA, USA: AMTA. Retrieved from http://amta2012.amtaweb.org/AMTA2012Files/start.htm

Krings, H. P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, OH, USA: The Kent State University Press.

Kruger, J. L. (2012). Making meaning in AVT: Eye tracking and viewer construction of narrative. *Perspectives: Studies in Translatology*, *20*(1), 67-86.

Lacruz, I., Denkowski, M., & Lavie, A. (2014). Cognitive Demand and Cognitive Effort in Post-Editing [online]. In *Proceedings of the 3rd Workshop on Post-Editing Technology and Practice (WPTP3)* (pp. 73-84). San Diego, CA, USA: AMTA. Retrieved from http://amtaweb.org/AMTA2014Proceedings/AMTA2014Proceedings_PEWorkshop_final.pdf

Lacruz, I., Shreve, G., & Angelone, E. (2012) Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)* (pp. 29-38). San Diego, CA, USA: AMTA. Retrieved from http://amta2012.amtaweb.org/AMTA2012Files/start.htm

Languages and the Media. (2004). *New markets, new tools. Post-conference report*. Berlin, Germany: ICWE. Retrieved from http://www.languages-media.com/downloads/postreport_2004.pdf

Lavie, A. & Agarwal, A. (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* (pp. 228-231). Prague, Czech Republic: Association for Computational Linguistics.

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

Lemmetty, S. (1999). *Review of Speech Synthesis Technology* (Unpublished MA thesis). Helsinki University of Technology, Finland. Retrieved from http://research.spa.aalto.fi/publications/theses/lemmetty_mst/thesis.pdf

Ljunglöf, P., Derbring, S., & Olsson, M. (2012). A free and open-source tool that reads movie subtitles aloud. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies* (pp. 1-4). Montréal, Canada: Association for Computational Linguistics (ACL).

Llisterri, J., Fernàndez, N., Gudayol, F., Poyatos, J. J., & Martí, J. (1993). Testing users' acceptance of Ciber232, a test to speech system used by blind people. In *Speech and Language Technology for Disabled Persons. Proceedings of an ESCA Workshop* (pp. 203-206). Stockholm, Sweden: KTH-ESCA.

López Vera, J. F. (2006). Translating Audio description scripts: The way forward? Tentative first stage project results. In *MuTra - Multidimensional Translation Conference Proceedings. Audiovisual Translation Scenarios* (no page numbers). Copenhagen, Denmark: MuTra. Retrieved from http://www.euroconferences.info/proceedings/2006_Proceedings/2006_proceedings.html

Luque, O. (2009). *El papel de la metáfora en la recepción de los GAD: una investigación empírica* (Unpublished MA thesis). Universidad de Granada, Spain.

Mączyńska, M. (2011). *TTS AD with audio subtitling to a non-fiction film. A case study based on La Soufriere by Werner Herzog* (Unpublished MA thesis). University of Warsaw, Poland.

Maszerowska, A., Matamala, A., & Orero, P. (Eds.). (2014). *Audio Description. New perspectives illustrated.* Amsterdam, The Netherlands: John Benjamins Publishing Company.

Matamala, A. (2005). Live audio description in Catalonia. *Translating Today, 4*, 9-11.

Matamala, A. (2006). La accesibilidad en los medios: aspectos lingüísticos y retos de formación. In R. Pérez-Amat & Á. Pérez-Ugena (Eds.), *Sociedad, integración y televisión en España* (pp. 293–306). Madrid, Spain: Laberinto Comunicación.

Matamala, A. (2015). The ALST project: Technologies for audiovisual translation. In *Proceedings of the 37th Conference Translating and the Computer* (pp. 79-89). Geneva, Switzerland: Editions Tradulex. Retrieved from https://ddd.uab.cat/pub/poncom/2015/144781/matamala_asling2015.pdf

Matamala, A., Fernández-Torné, A., & Ortiz-Boix, C. (2013). Enhancing sensorial and linguistic accessibility: further developments in the TECNACC and ALST projects. Presentation, 5th International Conference Media for All. Audiovisual Translation: Expanding Borders, Dubrovnik, Croatia. Retrieved from http://ddd.uab.cat/record/116868

Mazur, I., & Chmiel, A. (2012). Towards common European audio description guidelines: Results of the Pear Tree Project. *Perspectives: Studies in Translatology, 20*(1), 5-23. http://dx.doi.org/10.1080/0907676x.2011.632687

Mazur, I., & Chmiel, A. (2016). Should Audio Description Reflect the Way Sighted Viewers Look at Films? Combining Eye-Tracking and Reception Study Data. In A. Matamala & P. Orero (Eds.), *Researching Audio Description. New Approaches* (pp. 97-121). London, UK: Palgrave MacMillan. http://dx.doi.org/10.1057/978-1-137-56917-2_6

Melero, M., Oliver, A., & Badia, T. (2006). Automatic Multilingual Subtitling in the eTITLE project. In *Proceedings of Translating and the Computer: Vol. 28* (no page numbers). London, UK: ASLIB.

Mieskes, M., & Martínez Pérez, J. (2011). *A Web-based Editor for Audio-titling using Synthetic Speech*. Presentation, 3rd International Symposium on Live Subtitling with Speech Recognition, Antwerp, Belgium. Retrieved from http://www.respeaking.net/Antwerp%202011/Webbased_editor.pdf

Moorkens, J. & O'Brien, S. (2015). Post-editing evaluations: trade-offs between novice and profesional participant. In *Proceedings of the 18th annual conference of the European Association for Machine Translation (EAMT 2015)* (pp. 75-81). Antalya, Turkey: EAMT. Retrieved from https://aclweb.org/anthology/W/W15/W15-4910.pdf

Nichols, M. (Director) (2004). *Closer* [Film]. USA: Columbia Pictures.

Nielsen, S. & Bothe H. (2007). SUBPAL: A Device for Reading Aloud Subtitles from Television and Cinema. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments:*

Assistive Technology for All Ages (CVHI 2007). Granada, Spain: CEUR-WS. http://ceur-ws.org/Vol-415/paper17.pdf

O'Brien, S. (2004). Machine translatability and post-editing effort: How do they relate? *Translating and the Computer, 26*, no page numbers.

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

O'Brien, S. (2009). Eye tracking in translation process research: methodological challenges and solutions. In I. Mees, F. Alves & S. Göpferich, *Methodology, technology and innovation in translation process research: a tribute to Arnt Lykke Jakobsen* (pp. 251-266). Copenhagen, Denmark: Samfundslitteratur.

O'Brien, S. (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 AMTA 2010* (no page numbers). Denver, CO, USA: AMTA. Retrieved from <http://amta2010.amtaweb.org/AMTA/papers/6-01-ObrienPostEdit.pdf

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation, 25*, 197–215. Retrieved from http://doras.dcu.ie/17154/1/Towards_Predicting_Postediting_Productivity_Final_2.pdf

O'Hagan, M. (2003). Can language technology respond to the subtitler's dilemma? A preliminary study. In *Proceedings of Translating and the Computer: Vol. 25* (no page numbers). London, UK: ASLIB.

Oncins, E., Lopes, O., Orero, P., Serrano, J., & Carrabina J. (2013). All together now: a multi-language and multi-system mobile application to make living performing arts accessible. *Jostrans*, *20*, 147-164.

Orero, P. (2007a). Audio subtitling: A possible solution for opera accessibility in Catalonia. *TradTerm*, *13*, 135-149.

Orero, P. (2007b). Sampling audio description in Europe. In J. Díaz Cintas, P. Orero, & A. Remael (Eds.), *Media for All. Subtitling for the Deaf, Audio Description, and Sign Language* (pp. 111-125). Amsterdam/New York: Rodopi.

Orero, P. & Matamala, A. (2007). Accessible opera: overcoming linguistic and sensorial barriers. *Perspectives. Studies in Translatology*, *15*(4), 262-277. Retrieved from https://ddd.uab.cat/record/117149

Orero, P. & Vilaró, A. (2012). Eye tracking analysis of minor details in films for audio description. *MonTI. Monografías de Traducción e Interpretación*, *4*, 295-319.

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

Ortiz-Boix, C. & Matamala, A. (2015). Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: A Three-Level Approach. In *Proceedings of the 4th Workshop on Post-Editing Technology and Practice (WPTP4)* (pp. 2-15). Miami, FL, USA: AMTA. Retrieved from <http://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf>.

Ortiz-Boix, C. & Matamala, A. (Forthcoming). Post-editing Wildlife Documentary Films: A New Possible Scenario? *Jostrans*, *26*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pazos, P. (2012). *Audiosubtitulación: Una posible solución para la accesibilidad a los medios audiovisuales* (Unpublished MA thesis). Universitat Autònoma de Barcelona, Spain.

Peli, E., Fine, E., & Labianca, A. (1996). Evaluating Visual Information Provided by Audio Description. *Journal Of Visual Impairment & Blindness*, *90*(5), 378-385.

Piperidis, S., Demiros, I., & Prokopidis, P. (2004). *Multimodal Multilingual Information Processing for Automatic Subtitle Generation: Resources, Methods and System Architecture (MUSA).* Presentation, Languages and The Media, Berlin, Germany. Retrieved from http://sifnos.ilsp.gr/musa/LM/Languages and the Media-Berlin-Nov 2004-MUSA.ppt

Piperidis, S., Demiros, I., Prokopidis, P., Vanroose, P., Hoethker, A., Daelemans, W., Sklavounou, E., Konstantinou, M., & Karavidis, Y. (2004). Multimodal

multilingual resources in the subtitling process. In P*roceedings of the Fourth International Conference on Language Resources and Evaluation LREC-2004* (pp. 205-208). Paris, France: ELRA.

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

Popovic, M., Avramidis, E., Burchardt, A., Hunsicker, S., Schmeier, S., Tscherwinka, C., Vilar, D., & Uszkoreit, H. (2013). Learning from human judgments of machine translation output. In *Proceedings of the Machine Translation Summit XIV* (pp. 231-238). Nice, France: AMTA.

Popowich, F., McFetridge, P., Turcato, D., & Toole, J. (2000). Machine translation of closed captions. *Machine Translation, 15*(4), 311-341.

Rai, S. (2009). *Bollywood for all: The demand for audio described Bollywood films*. London, UK: Royal National Institute of Blind People (RNIB).

Ramos, M. (2013). *El impacto emocional de la audiodescripción* (PhD Thesis). Universidad de Murcia, Spain.

Remael, A. (2012). Audio-description with audio-subtitling: Making multilingual Dutch films: Manipulating Textual Cohesion on Different Levels. *Meta: Journal des Traducteurs / Meta: Translator's Journal*, *57*(2), 385. http://dx.doi.org/10.7202/1013952ar

Remael, A. & Vercauteren, G. (2007). Audio describing the exposition phase of films. Teaching students what to choose. *TRANS*, *11*, 73-93.

Remael, A. & Vercauteren, G. (2010). The translation of recorded audio description from English into Dutch. *Perspectives: Studies in Translatology, 18*(3), 155–171.

Remael, A., Orero, P., & Carroll, M. (2012). *Audiovisual Translation and Media Accessibility at the Crossroads: Media for All 3*. Amsterdam, New York, Kenilworth: Rodopi.

Remael, A., Reviers, N., & Vercauteren, G. (2015). *Pictures painted in words: ADLAB Audio Description Guidelines*. Trieste, Italy: EUT Edizioni Università di Trieste.

Rodríguez Posadas, G. & Sánchez Agudo, C. (2007). *Traducción de guiones audiodescriptivos: doble traducción, doble traición*. Presentation, AMADIS '07 Congress of the Centro Español de Subtitulado y Audiodescripción (CESyA), Granada, Spain.

Romero-Fresco, P. & Fryer, L. (2013). Could audio-described films benefit from audio introductions? An audience response study. *Journal of Visual Impairment & Blindness, 107*(4), 287-295.

Roturier, J., Mitchell, L., & Silva, D. (2013). The ACCEPT post-editing environment: a flexible and customisable online tool to perform and analyse machine translation post-editing. In *Proceedings of the 2nd Workshop on Post-Editing Technology and Practice (WPTP2)* (pp. 119-128). Nice, France: AMTA. Retrieved from http://www.mt-archive.info/10/MTS-2013-W2-Roturier.pdf

Saldanha, G., & O'Brien, S. (2013). *Research Methodologies in Translation Studies*. Manchester, UK: St. Jerome Publishing.

Salway, A. (2004). *AuDesc system specification and prototypes. TIWO: Television in Words)*. Guildford, UK: University of Surrey. Retrieved from http://www.bbrel.co.uk/pdfs/TIWO_Television_in_Words_Deliverable_3.pdf

Salway, A., Tomadaki, E., & Vassiliou, A. (2004). *Building and analysing a corpus of AD scripts. TIWO Television in Words. Report on Worckpackage 2*. Guildford, UK: University of Surrey. Retrieved from http://www.bbrel.co.uk/pdfs/TIWO_Television_in_Words_Deliverable_2.pdf

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (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* (pp. 223-231). Cambridge, MA, USA: AMTA.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation* (pp. 73-80). Leuven, Belgium: EAMT. Retrieved from http://www.mt-archive.info/EAMT-2011-Specia.pdf

Suojanen, T., Koskinen, K., & Tuominen, T. (2015). *User-centered translation*. London, UK: Routledge, Taylor & Francis Group.

Szarkowska, A. (2011). Text-to-speech audio description: towards wider availability of AD. *The Journal of Specialised Translation*, *15*, 142-269.

Szarkowska, A., & Jankowska, A. (2012). Text-to-speech audio description of voice-over films. A case study of audio described Volver in Polish. In E. Perego (Ed.), *Emerging topics in translation: Audio description* (pp. 81-94). Trieste, Italy: EUT Edizioni Università di Trieste.

Szarkowska, A., & Jankowska, A. (2015). Audio describing foreign films. *The Journal of Specialised Translation*, *23*, 243-162.

Tatsumi, M. & Roturier, J. (2010). Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship? In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry* (pp. 43-51). Denver, CO, USA: AMTA. Retrieved from http://www.mt-archive.info/JEC-2010-Tatsumi.pdf

TAUS & CNGL. (2010). *Machine translation postediting guidelines*. De Rijp, The Netherlands: TAUS. Retrieved from http://taus-website-media.s3.amazonaws.com/images/stories/guidelines/taus-cngl-machine-translation-postediting-guidelines.pdf

Temizöz, Ö. (2012). Machine translation and postediting. In *The European Society for Translation Studies State-of-the-Art Research Report*. Herentals, Belgium: European Society for Translation Studies.

Udo, J. P., & Fels, D. I. (2009). Suit the action to the word, the word to the action: An unconventional approach to describing Shakespeare's Hamlet. *Journal of Visual Impairment & Blindness, 103*(3), 178-183.

Vázquez, Y. & Huckvale, M. (2002). The reliability of the ITU-t p.85 standard for the evaluation of text-to-speech systems. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002* (pp. 329-332). Denver, CO, USA: ISCA.

Verboom, M., Crombie, D., Dijk, E., & Theunisz, M. (2002). Spoken Subtitles: Making Subtitled TV Programmes Accessible. In *Proceedings of the 8th International Conference Computers Helping People with Special Needs ICCHP 2002* (pp. 295-302). Berlin & Heidelberg, Germany: Springer Verlag. http://dx.doi.org/10.1007/3-540-45491-8_62

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language, 19*(1), 55-83. http://dx.doi.org/10.1016/j.csl.2003.12.001

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

Walczak, A. (2010). *Audio description for children. A case study of text-to-speech audio description of educational animation series* Once Upon a Time... Life (Unpublished MA Thesis). University of Warsaw, Poland.

Walczak, A., & Szarkowska, A. (2012). Text-to-speech audio description to educational materials for visually-impaired children. In S. Bruti & E. Di Giovanni (Eds.), *Audio Visual Translation across Europe: An ever-changing landscape* (pp. 209-234). Bern, Switzerland: Peter Lang.

Weaver, S. L. (2013). *Lifting the curtain on opera translation and accessibility: Translating opera for audiences with varying sensory ability* (Unpublished MA thesis). Durham University, UK.

World Health Organisation Media Centre. (2014). *Visual impairment and blindness. Fact Sheet N°282*. WHO. Retrieved from Visual impairment and blindness. (2014). World Health Organization. Retrieved 23 May 2016, from http://www.who.int/mediacentre/factsheets/fs282/en/