2175-7968.2015v35n2p308

SPEAKER DIARIZATION AND SPEECH RECOGNITION IN THE SEMI-AUTOMATIZATION OF AUDIO DESCRIPTION: AN EXPLORATORY STUDY ON FUTURE POSSIBILITIES?

Héctor Delgado[*]
Universitat Autònoma de Barcelona

Anna Matamala[**]
Universitat Autònoma de Barcelona

Javier Serrano[***]
Universitat Autònoma de Barcelona

[*] BS in Computer Science Engineering by Universidad de Sevilla, Spain, and MS in Multimedia Technologies by Universitat Autònoma de Barcelona, Spain. PhD candidate at the Department of Telecommunications and Systems Engineering at Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain. E-mail: hecdelflo@gmail.com

[**] BA in Translation and Interpreting by Universitat Autònoma de Barcelona, and PhD in Applied Linguistics by Universitat Pompeu Fabra (Barcelona). Tenured senior lecturer at Universitat Autònoma de Barcelona (Spain). E-mail: anna.matamala@uab.cat

[***] BA in Computer Science (Universitat Autònoma de Barcelona) and PhD in Automatic Control (Computer Science Program, UAB). Associate Professor at Universitat Autònoma de Barcelona. E-email: javier.serrano@uab.cat

Abstract: This article presents an overview of the technological components used in the process of audio description, and suggests a new scenario in which speech recognition, machine translation, and text-to-speech, with the corresponding human revision, could be used to increase audio description provision. The article focuses on a process in which both speaker diarization and speech recognition are used in order to obtain a semi-automatic transcription of the audio description track. The technical process is presented and experimental results are summarized.

Keywords: Audio description. Accessibility. Speaker diarization. Speech recognition. Technology.

DIARIZACIÓN Y RECONOCIMIENTO DE HABLA EN LA SEMIAUTOMATIZACIÓN DE LA AUDIODESCRIPCIÓN: UN ESTUDIO EXPLORATORIO SOBRE POSIBILIDADES FUTURAS

Resumem: Este artículo presenta una visión panorámica de los componentes tecnológicos usados en el proceso de audiodescripción y propone un nuevo escenario en el que se aplicarían el reconocimiento de habla, la traducción automática y la síntesis de habla, con su correspondiente revisión humana, para incrementar la cantidad de audiodescripciones disponibles. El artículo describe un proceso en el que la diarización y el reconocimiento de habla permiten obtener una transcripción semiautomática de la audiodescripción. El artículo presenta detalladamente el proceso técnico así como un resumen de los resultados experimentales.- In a second language.

Palabras clave: Audiodescripción. Accesibilidad. Diarización. Reconocimiento de habla. Tecnología.

Audio description (AD) is an access service providing an auditory translation of visual content, primarily aimed at those who are blind and visually impaired. The provision of audio descriptions is uneven across types of content, platforms and countries where the content is delivered (ADLAB, 2012). Whilst in certain countries audio description is still an emerging access service, in others audio description provision is rapidly increasing due to higher awareness and regulation.

However, there is still a lot of content that is not audio described, a situation which is especially striking in user-generated content on social media. Software and technology systems to facilitate what could be called “fan audio descriptions” have been put forward (see, for instance, YouDescribe (youdescribe.org), MAGpie (ncam.wgbh.org/invent_build/web_multimedia/tools-guidelines/magpie) or LiveDescribe (imdc.ca/ourprojects/livedescribe)). But the number of amateur audio descriptions in social media is still relatively low.

It is often said that creating an AD is costly because it takes a lot of time and human effort. Departing from the DTV4ALL report (2009), one could say that two typical workflows are to be found in the process of creating AD:

a. Scenario 1: the describer is in charge of creating and voicing the AD, and there is usually another professional carrying out a quality control check.
b. Scenario 2: the describer (often working in teams involving a blind professional) creates an AD script and afterwards a voice talent records it.

In both scenarios ADs are generally created in only one language, for a certain audience that understands that language. In order to speed up the process of AD, reduce associated costs and offer wider availability, some solutions have been put forward: human translation has been one of the suggested alternatives to the process of creating audio descriptions (Matamala, 2006; Jankowska 2015) and text-to-speech systems have been researched as alternatives to human voicing (Szarkowska, 2011; Fernández-Torné & Matamala, forthcoming).

In this changing scenario, the Spanish-funded project ALST (Technologies for sensorial and linguistic accessibility) aims to research whether additional semi-automatic components could be added to the process of creating an AD. ALST imagines a scenario in which an oral input into language A would travel along a semiautomatized workflow and finally be delivered orally into language B, as shown in Figure 1. This workflow would include three key technological components: a speech transcription process preceded by a speaker diarization and a speaker segmentation process, as discussed in this paper (step 1). A machine translation (MT) component, in which a draft translation would be generated by an engine and would be revised by a human post-editor (step 2). And a speech synthesis component, in which a text-to-speech system would be voicing the AD translation (step 3). This workflow is defined as semi-automatic, our belief being that a post-editing or revision of the automatic outputs in steps 1 and 2 would generally be needed to reach high quality standards. Especially when taking into account the wide variability of acoustic conditions and language variation found in audio descriptions.

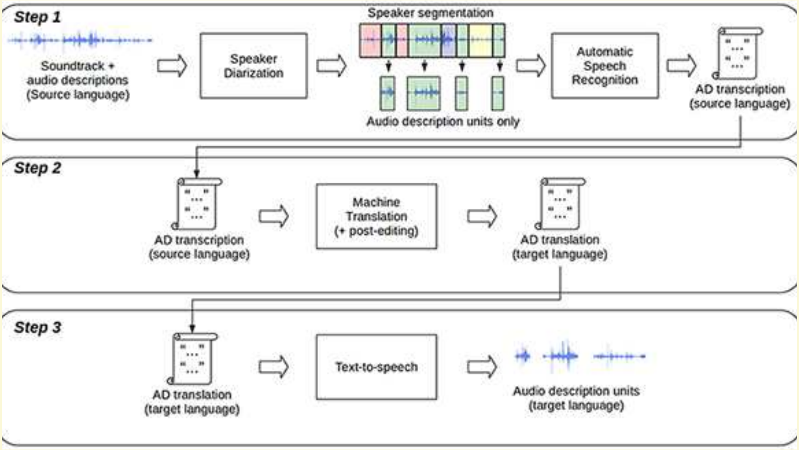
Step 1
Soundtrack + audio descriptions (Source language)
Speaker Diarization
Speaker segmentation
Audio description units only
Automatic Speech Recognition
AD transcription (source language)
Step 2
AD transcription (source language)
Machine Translation (+ post-editing)
AD translation (target language)
Step 3
AD translation (target language)
Text-to-speech
Audio description units (target language)

Figure 1. ALST proposed workflow

This semi-automatic process could have an impact in traditional workflows but could also be relevant in user-generated content and social media, where collaborative approaches supported by technology are usually welcome and quality expectations are different to a certain extent. It would be worth researching whether in these latter scenarios the process could be fully automatized and the human component included above (revision of speech recognition output and post-editing of machine translation output) could be removed.

With this framework in mind, the article begins by summarizing research in the field of AD that has tried to incorporate technological contents mainly in the process of creation but also, tangentially, in the process of delivering the AD. The second section describes the specific experiment that was developed to automatically transcribe an AD from a video content, and the experimental results. This is the focus of the article, and its aim is, on the one hand, to describe a technical process which semi-automatically extracts the AD track and generates a transcription, and, on the other, to describe preliminary testing results obtained with available data and technologies. Therefore, to present a fully operational system is not the aim of this article. Although one could argue that scripts could be obtained more easily from their authors or copyright owners, our research wanted to consider a scenario in which these transcripts are not easily available. This may be especially relevant in the era of social media and shared user-generated video content.

1. Technology in audio description

Two prototypical AD workflows have been described in the introduction above. The only technological component used for the creation of AD in these traditional scenarios is dedicated AD software such as Swift Adept (www.grassvalley.com/products/swift_adept) or Anglatècnic Fingertext (www.anglatecnic.cat/ca-40-Audiodescripció). This software usually facilitates the creation of time-coded audio descriptions and their recording and mixing.

However, research has gone a step further by investigating the application of other technologies in the process. Text-to-speech systems were the first ones to be researched, and have already attracted the attention of the industry. Szarkowska (2011) investigated what was termed “text-to-speech audio description” (TTS AD) in a wide project carried out in Poland. The reception of TTS AD was assessed in a monolingual feature film in Polish (Szarkowska, 2011), in a dubbed educational TV series for children (Walczak and Szarkowska, 2010), in a foreign fiction film with voice-over (Szarkowska and Jankowska, 2012), in a non-fiction film with audio subtitling (Macynska, 2011), and in a dubbed feature film (Drozdz-Kubik, 2011). Results show that most participants accept TTS AD, although it is not always the preferred solution.

Similarly, Kobayashi et al. (2009) describe the application of TTS AD in online videos on websites. A wide survey in which three kinds of voices were tested (human, standard TTS and prototype TTS) followed by in-depth interview sessions, both in Japan and in the US, showed that TTS AD was accepted by end-users, especially for relatively short videos and informational content.

Within the ALST project, Fernández-Torné and Matamala (forthcoming) present the results of an experiment in which 67 blind and visually impaired people assessed two synthetic voices when applied to audio description, as compared to two natural voices, also with promising results in terms of TTS AD acceptance.

In the industry, the company Swiss TXT is already planning to offer TTS AD (Caruso 2012), while a web-based editor for transforming text into speech has been developed by Mieskes and Martínez (2011).

Although text-to-speech has so far been the main technological component to be researched in AD, machine translation has been the focus of recent investigations, inspired by previous debates on the inclusion of human translation in the AD process. Indeed, human translation has been discussed in the literature as a possible solution to speed-up the process of AD creation with opposing views: whilst Matamala (2006), Bourne and Jiménez (2007) and Jankowska (2015) defend this scenario, Hyks (2005) considers it would take longer to translate an AD script than to create one. Vallverdú, in Matamala and Orero (2009), states that a human translation would involve transcribing the text, time-coding it and adapting it, all of which may result in a time-consuming task. A specific analysis based on two case studies is carried out by Remael and Vercauteren (2010), who highlight specific challenges linked to AD translation while acknowledging it would be a time-saving practice. However, no empirical evidence is provided in either of the previous papers. Fernández-Torné, Matamala and Ortiz-Boix (2012) go a step further and research the application of machine translation in the AD workflow within the ALST project.

More specifically, Ortiz-Boix and Matamala (forthcoming) investigate the application of MT in the field of AD by focusing on two closely related languages (Spanish-Catalan). The analysis uses two, free online engines and relies on a human assessment based on error categorization. Results show the feasibility of applying MT to filmic audio description in closely related languages. Dealing with another language pair (English into Catalan), Fernández-Torné and Matamala (2014) describe an experiment on machine translated AD. Participants in their experiments post-edited the machine translation output of five free online engines for a 3-minute AD clip containing 14 AD units. They also assessed these units in terms of post-editing necessity, post-editing difficulty, MT adequacy and MT fluency. A final ranking task completed this human assessment, which was complemented with automatic measures such as HBLEU, PE time and HTER. Results allowed a selection of the "best" machine translation engine, which was then used in an experiment where three scenarios were compared: post-editing a machine translated audio description excerpt, translating an audio description excerpt and creating an audio description ex novo. 15 participants took part in this experiment and results in terms of temporal effort (time spent on the task), technical effort (keyboard usage) and cognitive effort (measured in terms of pause to word ratio and average pause ratio) were obtained.

Leaving aside the process of creation and briefly getting in to the field of delivery, specific systems and mobile applications have also been developed to offer easier access to AD. At Universitat Autònoma de Barcelona (UAB), Oncins et al. (2013) have developed the Universal Accessibility System (UAS), a multi-language and multi-system mobile application used to make live events accessible. UAS is designed to offer automatic AD through TTS as well as other features. On professional environments, AudescMobile and ArtAcces are apps that provide both subtitling and audio description in a similar way. Additionally, a trial by RNIB is underway to test the delivery of audio descriptions via the MovieReading mobile app (https://www.rnib.org.uk/audio-description-app). These apps generally identify the exact point in the film/TV programme and synch a downloaded AD track, which for the moment has been created by a human describer, but could easily incorporate text-to-speech in the future.

2. The speech recognition experiment: technical process

Although the ALST project is wider in scope, this paper focuses on the description of an experiment aiming to extract and transcribe the audio description from a movie soundtrack. The semi-automatization of this task could prove useful in a scenario in which no written scripts are available. As indicated previously, a written script may be required in a semi-automatic process of audio description, in which a draft audio description could be generated by a machine translation engine later revised by a human post-editor.

The movie selected for analysis was *Closer* (Mike Nichols, 2004), both in its original version in English and in its Catalan-dubbed version. The selection was due to the fact that the same movie had previously been used in complementary technological experiments in the ALST project (machine translation, text-to-speech), and a single input was preferred. The initial selection of this movie was mainly motivated by the availability of all necessary material (written time-coded transcripts and video/audio tracks) to carry out the quality assessment.

The technical process for the automatic audio description extraction and transcription contained essentially the following tasks: soundtrack extraction, speech activity detection, speaker diarization, and finally, speech-to-text transcription.

First of all, the movie soundtrack was extracted from the video file and then converted into a suitable format. More specifically, the two available audio channels (from the original stereo sound file) were mixed together into a single mono channel. Then, downsampling was carried out in order to obtain a 16 KHz, 16-bit, PCM wave file. The resulting audio file contained the sounds of the movie plus the AD mixed together.

Secondly, an audio segmentation was performed to the wave file. This process aimed to remove all non-speech content from the audio, and only keep speech content. This process is usually referred to as Speech Activity Detection (SAD) or Voice Activity Detection (VAD), and it is a very common pre-processing tool for other speech-related tasks. In this experiment, SAD was done in order to provide a speech signal as clean as possible to the next module in the processing chain. This process was carried out with the acoustic segmentation tool included in the ALIZE toolkit developed by the University of Avignon, and described in Fredouille et al. (2009).

Thirdly, the AD units within the audio track had to be extracted. When there is prior information and available data about the describer speaking, a speaker model trained on the describer's voice could have been used to extract the AD turns by means of the technique called "speaker tracking". However, the proposed scenario of application assumes that no training data is available. In such a situation, the only option is to turn to unsupervised approaches which do not depend on training data. In this regard, the task called "speaker diarization" aims to segment a speech stream into speaker-homogeneous segments (i.e. each segment contains speech from a single speaker), assigning them a unique abstract identifier, according to the speaker identities. This process is performed without any prior information about the participating speakers or their number, thus, speaker diarization seems to meet the needs of the proposed scenario. As a result, speaker diarization was performed over the speech signal output by the SAD module. The result of this process was a text file containing information about the detected speaker-homogeneous segments. For every segment, this information comprises a speaker ID, beginning time-code, and ending time-code. The speaker diarization system used at this stage is based on the Binary Key speaker modeling (Delgado et al., 2014).

The speaker diarization system detected a number of different speakers within the audio stream, assigning them a unique abstract identifier. The fourth step consisted of identifying the abstract ID which corresponded to the describer. Once again, this process could be done automatically through speaker recognition, but since there was no prior data, automatic speaker recognition could not be performed and, consequently, the selection had to be done manually. This was the only step that required human supervision.

After the diarization stage, some processing was applied to the obtained segments: very short segments, less than 1 second long, were discarded. Close segments with a separation smaller than 1 second were merged together. And an increase of 0.5 seconds both at the beginning and at the end was applied to all segments, in order to add a period of silence before and after the describer turns. This process was carried out as it is usually beneficial for automatic speech recognition. Finally, the resulting segments were used to split the signal into audio description units, and the rest of the speech was discarded. Each audio description unit was isolated in an individual wave file.

Finally, the resulting AD sound files were transcribed, only in its original English version, using two automatic speech recognition systems (ASR). The first system used was a large vocabulary continuous speech transcription system – tailored to achieve quality transcriptions of broadcast news audio, and trained on large amounts of broadcast news audio and text (system A). The second system was a commercial dictation system trained for single speaker dictation purposes (system B).

3. Experimental results

This section shows the obtained experimental results. Speaker diarization and ASR modules are evaluated separately.

As for the speaker diarization, performance was measured in terms of Diarization Error Rate (DER). DER is the most common metric to assess speaker diarization quality. DER is the sum of three different sources of error. First, the miss speech time category is the percentage of speech present in the reference that the system has not been able to detect. Second, the false alarm speech time is the percentage of speech detected by the system which is not actually labeled in the reference. Finally, the speaker error time is the percentage of time that the system has assigned to an incorrect speaker. Computing DER requires a text file containing the reference speaker turns of the movie being processed.

In this experiment, the complete speaker turns reference was not available. Only the information about the describer turns was provided. As such, an actual evaluation of speaker diarization performance cannot be performed. Instead, an alternative evaluation was performed, taking into account only the describer segments. At that point, all segments not belonging to the describer were removed from the diarization output, and DER was computed by comparing only the describer segments obtained by the diarization systems against the reference segments.

Since there is only one speaker in the reference, speaker errors cannot occur, and speaker error time is equal to zero. Therefore, here DER is the sum of false alarm errors and miss speech errors. The obtained results are shown in the table below:

	miss	false alarm	speaker error	DER
Catalan	18.7%	3.9%	0%	22.6%
English	11.8%	9.2%	0%	21.03%

Table 1. DER for speaker diarization

It can be observed that the predominant error in both tests is the miss speech time. A possible explanation for this is that with this type of audio (movie soundtrack), there is a high sound variability: speakers can talk under a wide range of acoustic conditions, such as over music, background noises, or even other voices. This variability may lead the speaker diarization system to generate more than one cluster for each speaker. When the majority cluster is manually selected as the describer cluster, audio description regions assigned to other clusters are systematically lost. With regard to top false alarm error, a greater difference is appreciated between Catalan and English tests. In the case of Catalan, only 3.9% of the time is assigned to regions where the describer is not participating. However, in the English experiment, a higher value of 9.2% is obtained.

Each type of error has a different impact on our workflow: regions classified as miss speech time are not present in the resulting audio files. Consequently, those segments corresponding to the describer will not be processed by the ASR system. On the other hand, regions of false alarm speech are included in the resulting audio, meaning that speech from other speakers will be introduced in the ASR system.

As for the automatic speech recognition results, the Word Error Rate metric was used, as it is the most extended metric to assess speech recognition performance. WER is defined as the sum of all possible errors divided by the actual number of words in the reference. These errors include insertions, deletions, and substitutions.

Table 2 summarizes the results obtained in English, where ASR systems were available for preliminary testing.

	WER	Hits	Deletions	Substitutions	Insertions	#words
System 1	64.43	2427	1086	3310	996	6823
System 2	47.18	3604	1748	1471	458	6823

Table 2. ASR results

Performance of the systems was not very high, and this is mainly due to the mismatch between the training conditions of the various systems and the employed AD test materials. As already discussed, the selected systems were intended for transcribing broadcast news or for single speaker dictation purposes and as a result achieve higher levels of performances in these domains. The English transcription system 1 reaches WER between 15-20% in broadcast news content (Álvarez et al., 2015) and the English dictation system 2 quotes accuracy rates above 90% when single speaker clean quality audios are employed. These results are state-of-the-art in the corresponding domains.

Overall, it seems that for existing ASR systems to achieve better performance on AD material, they need to be adapted to the specific acoustic and vocabulary conditions of the AD content. Acoustically, AD audio contains vocal music and/or dialogues in the background. Linguistically, each AD track uses content-specific vocabulary. Future experiments on automatic transcription of AD tracks should take this into account in order to achieve usable transcriptions.

4. Conclusions and further work

Research in the field of audiovisual translation and media accessibility has traditionally focused on descriptive approaches which have later been complemented with experimental research on user reception. Technological research is relatively recent, and both fully functional proposals and experimental approaches suggesting new concepts should have a place in this new research scenario. This paper presents an instance of such research in which audiovisual translators, media access experts and engineers have cooperated towards developing and testing a new experimental workflow. Preliminary results show a low performance of the speech recognition systems which may impede a professional application of this technical process at this stage, but room for further research in this field is obvious. Using trained speech recognition systems and widening the experimental data would be the first natural extension of this investigation and would facilitate obtaining stronger results. Further research could encompass other languages, and the ultimate aim would be to integrate this first semi-automatic task with additional technical components in the chain (machine translation / text-to-speech) in order to semi-automatize the process of AD creation. Assessing the need for a human revision at various stages in the worflow would also be a relevant research topic, as well as dealing with delivery mechanisms and interoperability aspects.

Research possibilities in the field of AD and technologies are manifold, but this paper can be considered a first step in applying speech recognition and speaker diarization in the field as a new way to semi-automatize the process of AD creation.

Acknowledgements

This research has been funded by the Spanish Ministry of Economy as part of the ALST (Technologies for sensorial and linguistic accessibility) project (FFI-2012-31024), starting in 2013 and ending in 2015. It is also part of TransMedia Catalonia research (Catalan Government funds, 2014SGR27).

References

ADLAB (2012). *Report on user needs assessment. Report no. 1, ADLAB (Audio Description: Lifelong Access to the Blind) project*. Retrieved from www.adlabproject.eu.

Álvarez, A.; Mendes, C.; Raffaello, M.; Luis, T.; Paulo, S.; Piccinini, N.; Arzelus, H.; Neto, J.; Aliprandi, C., & Del Pozo, A. (2015). *Automating live and batchsubtitling of multimedia contents for several European languages*. *Multimedia Tools and Applications* (MTAP).

Bourne, J., & Jiménez, C. (2007). From the visual to the verbal in two languages: a contrastive analysis of the audio description of *The Hours* in English and Spanish. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All. Subtitling for the Deaf, Audio Description, and Sign Languages* (pp. 175-188). Amsterdam: Rodopi.

Caruso, B. (2012). Audio description using speech synthesis. In *Languages and the Media. 9th International Conference on Language Transfer in Audiovisual Media. Conference Catalogue* (pp. 59-60). Berlin: ICWE.

Delgado, H., Fredouille, C., & Serrano, J. (2014). Towards a complete binary key system for the speaker diarization task. *Interspeech 2014. Proceedings of the 15th Annual Conference of the International Speech Communication Association* (pp. 572-576). Singapore.

Drożdż-Kubik, J. (2011). *Harry Potter iKamieńFilozoficznysłowemmalowany – czylibadanieodbiorufilmu* z *audiodeskrypcją z synteząmowy*. Unpublished MA Thesis. Krakow: Jagiellonian University, Poland.

DTV4ALL (2009). *Digital Television for All. D2.3. Interim Report on Pilot services*. Retrieved from http://dea.brunel.ac.uk/dtv4all/ICT-PSP-224994-D23.pdf

Fernández-Torné, A., & Matamala, A. (2014, November). *Machine translation and audio description. Is it worth it? Assessing the post-editing effort*. Paper presented at Languages and the Media. 10th International Conference on Languages Transfer in Audiovisual Media, Berlin, Germany.

Fernández-Torné, A., & Matamala, A. (forthcoming). Text-to-speech vs human voiced audio descriptions: a reception study in films dubbed into Catalan. *Jostrans. The Journal of Specialised Translation.*

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

Fredouille, C.; Bozonnet, S., & Evans, N.W.D. (2009) The LIA- EURECOM RT‘09 Speaker Diarization System.*RT’09, NIST Rich Transcription Workshop*. Florida, USA. Retrieved from http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/LIA-EURECOM_paper.pdf

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

Jankowska, A. (2015). *Translating audio description scripts. Translating as a new strategy of creating audio description.* Frankfurt: Peter Lang.

Kobayashi, M., Fukuda, K., Takagi, H., & Asakawa, C. (2009). Providing synthesized audio description for online videos. *ASSETS ’09: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*(pp. 249-250). New York, USA: ACM.

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. Warsaw: University of Warsaw, Poland.

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: Laberinto.

Matamala, A., & Orero, P. (2009). L’accessibilitat a Televisió de Catalunya: parlemamb Rosa Vallverdú, directora del departament de Subtitulació de TVC. *Quaderns, Revista de Traducció*, *16*, 301-312.

Mieskes, M., & Martínez Pérez, J. (2011). *A web-based editor for audio-titling using synthetic speech*. Paper presented at the 3rd International Symposium on Live Subtitling with Speech Recognition, Antwerp, Belgium. Retrieved from http://www.respeaking.net/Antwerp%202011/Webbased_editor.pdf

Moreno, A.; Febrer, A., & Márquez, L. (2006). Generation of Language Resources for the Development of Speech Technologies in Catalan. *Proceedings of* *the Language Resources and Evaluation Conference LREC 06* (pp. 1632-1635). LREC: Genoa, Italy.

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. The Journal of Specialised Translation, 20*, 147-164.

Ortiz-Boix, C., & Matamala, A. (forthcoming). Accessibility and multilingualism: an exploratory study on the machine translation of audio descriptions. *Trans. Revista de Traductología.*

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

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

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.) (2012). *Emerging topics in translation: Audio description* (pp. 81-98). Trieste, Italy: Edizioni Università di Trieste.

Walczak, A., & Szarkowska, A. (2012). Text-to-speech audio description of 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/Berlin: Peter Lang.

Recebido em: 05/06/2015
Aceito em: 04/08/2015