Building an Audio Description Multilingual Multimodal Corpus: The VIW Project

Anna Matamala, Marta Villegas
Universitat Autònoma de Barcelona
Edifici K, 08193 Bellaterra
E-mail: anna.matamala@uab.cat, marta.villegas@uab.cat

Abstract

This paper presents an audio description multilingual and multimodal corpus developed within the VIW (Visual Into Words) Project. A short fiction film was created in English for the project and was dubbed into Spanish and Catalan. Then, 10 audio descriptions in Catalan, 10 in English and 10 in Spanish were commissioned to professional describers. All these data were annotated at two levels (cinematic and linguistic) and were analysed using ELAN. The corpus is an innovative tool in the field of audiovisual translation research which allows for comparative analyses both intralingually and interlingually. Examples of possible analyses are put forward in the paper.

Keywords: audio description, audiovisual translation, multimodal corpus, multilingual corpus

1. Introduction

Audio description (AD) is an intersemiotic translation in which images are translated into words (Maszerowska et al., 2014). These words are delivered aurally to an audience that does not have access to the visuals, mainly the blind and visually impaired but also other users who for various reasons cannot access the visual content. In audiovisual productions, AD is interspersed in the segments where no dialogue and no relevant sounds are heard. Its aim is that the audience can understand and enjoy the audiovisual content only through the audio channel. One could say that AD has been provided informally by sighted people who, for instance, watch television with blind or visually impaired friends or family. Volunteers have also played a key role in making many cultural activities accessible to all. However, AD as a professional access service is more recent, and still non-existent in certain countries (Orero, 2007), despite the fact that accessibility has been included as a human right in the UN Convention on the Rights of Persons with Disabilities and there is increasing legislation promoting it (see the recent EU proposal for an Accessibility Act).

Guidelines have been developed by standardization bodies, regulators and associations (Matamala and Orero, 2013) to help describers in the complex task of translating all the nuances provided by the images into a limited number of sentences or words. A set of strategies designed within the ADLAB project (Remael et al., 2016) are a useful tool to help describers in their choices.

Research on AD is also recent and has been integrated within the frame of audiovisual (AV) translation studies (Braun 2008). Investigations on AD have been mainly descriptive, dealing with specific practices such as theatre opera, art, and cinema. Case-studies have approached the analysis of various features in ADs, such as cultural references (Mangiron and Maszerowska, 2014), sometimes adopting a contrastive approach (Bourne and Jiménez, 2007). More recently, reception research with end users and technological aspects have been tackled. However, AD corpus research has been scarce, and there is still a lot to be learnt concerning both the process of AD and the final product.

This paper presents a corpus of ADs developed within a one-year project (Visuals Into Words, VIW), running from October 2015 until September 2016 under the Spanish Government *Europa Excelencia* funding scheme. VIW's ultimate aim is to create an open access platform that will allow for comparative research on AD, both intralingually an interlingually. To contextualise this project within AV translation studies research, Section 2 summarises the state of the art in corpus research in AD. Section 3 explains the project rationale. Section 4 describes the corpus in its current stage of development, and section 5 defines the corpus annotation procedures. Section 6 puts forward possible corpus exploitations.

2. Previous Work

Most AD research has focused on one film, sometimes expanding the corpus to a few films (Piety, 2004). Two relevant exceptions to this trend are TIWO and TRACCE.

TIWO (Television in Words) was a project led by Andrew Salway between 2002-2005 at the University of Surrey (UK) which aimed "to develop a computational understanding of storytelling in multimedia contexts, with a focus on the processes of AD" (Salway, 2007: 153). In order to do so, 91 audio description scripts in British English from three major producers of AD were collected, making up a corpus of 618,859 words (Salway, 2007: 155). The TIWO corpus allowed Salway to carry out a thorough analysis of the language of AD in English (Salway, 2007). It also compelled him to propose some ideas on assisted audio description, and to suggest how AD could be used for keyword-based video indexing.

On the other hand, TRACCE (Jiménez Hurtado et al., 2010) was a project led by Catalina Jiménez Hurtado between 2006 and 2009 at the University of Granada (Spain). A corpus of 300 films audio described in Spanish, plus 50 films in German, English and French, were collected. Most of the Spanish scripts came from the film archives of the Spanish blind association ONCE because at the time the corpus was created few commercially available audio described films were available in Spanish. A multimodal annotation system and a specific tool were developed (*Taggetti*) to tag the AD scripts (Jiménez Hurtado and Seibel 2012: 412). Annotations were created at three different levels: film narrative, camera language, and recurrent grammatical structures in the ADs. The tagging process was carried out manually in one-minute film segments called Meaning Units, which were composed of the AD script and the associated AV content.

Despite the relevance of both projects in AD corpus research, they are not freely available on the Internet, probably due to copyright issues. This is similar to what happens often in other fields of AV translation, where corpora have been created but have faced copyright constraints (Baños, Bruti and Zanotti, 2013).

On the other hand, sometimes a significant amount of data have been gathered, but they have not been incorporated into a systematic corpus. This is the case, for instance, of the Pear Tree project (Mazur and Kruger, 2012) developed within the DTV4ALL project. Participants from different countries collected descriptions of the same film, a clip created for Chafe's (1980) Pear Stories project that contained no dialogue, in order to identify cultural similarities and divergences.

3. Motivation

It is in this context that VIW was born. Inspired by Chafe's (1980) project, and its posterior implementation in AD (Mazur and Kruger, 2012), VIW aims to develop a multimodal and multilingual corpus of AD departing from a single stimulus, a short film created *ad hoc* in English, and translated into other languages. This corpus will allow studies to be carried out comparing the AD versions produced for one language but also contrasting various languages.

The project is built upon two pillars: on the one hand, it has a strong open access component. All materials will be freely available to the research community, through an open platform that is currently being designed (Creative Commons licence CC-BY-NC-SA). Copyright has been secured through agreements developed specifically for the project, both for the film (in English and in its translated versions) and the ADs created. On the other hand, it aims to be a scalable and expanding project. This means that, although very limited in size in its initial stages, the project is being designed and developed so that it can easily incorporate other languages and inputs provided by external researchers. This will be feasible thanks to a clear documentation of all the processes and the implementation of open access tools and licences.

4. Corpus description

This section describes the corpus considering its current stage of development, but also indicating further developments that will be achieved on its completion. It differentiates between the short film that is at the core of the project and the AD that have been created.

4.1 The Short Film

The short film was commissioned to a film director and produced specifically for the project. To make sure the film would be useful for AD research purposes, a literature review and experts' discussion allowed identification of the key elements that are considered challenges in AD. These included: characters and action, including gestures and facial expressions, spatial-temporal settings, film language, sound effects and silence, text on screen, and intertextual references (Maszerowska et al., 2015). The film director was instructed to create a short film with a standard narrative structure, various actions, and at least four characters speaking in English except for one, who would speak another language at least at some point so that subtitles could be added. Further instructions were to include at least three different spatial-temporal settings, and to incorporate some text on screen as well as opening and end credits. The director was told to include in the film narrative at least one sound that could not be easily identifiable, and to show silent passages for artistic purposes. Finally, the film director was made aware that the film would be audio described, hence segments without speech were needed to add the audio description.

It was considered that the film should last a minimum of 10 minutes to allow for research on user engagement, a hot topic in the AD research agenda. At the same time, it was considered that a much longer film would make more difficult its re-usage in experimental settings and, last but not least, it would also be difficult to support financially. This is why the film director was instructed to create a film between 12 and 15 minutes long. The result is the film "What happens while---", directed by Núria Nia, which lasts 14 minutes, and deals with how different characters envisage time.

Since our aim was to include AD in English, Catalan, and Spanish, a dubbed version of the short film was commissioned to a Barcelona-based dubbing studio. The same translator, dubbing director, and voice talents were used to create both the Catalan and the Spanish version. All three versions are available from the project website (http://pagines.uab.cat/viw).

4.2 The Audio Descriptions

Ten English AD, ten Spanish AD, and ten Catalan AD were commissioned to professional AD providers. They were requested to generate an AD of the short film following the usual professional standards. They were instructed to send an .mp4 file containing the final audio-video mix plus a time-coded script, without further specifications, over a period of approximately two weeks. Some providers offered the researcher the possibility to make changes to the AD, but it was decided not to intervene in the process and just accept the output as delivered.

As of March 2015, 10 versions per language are available. An experiment is also ongoing to gather ADs created by AD students to complement the current corpus. This would allow for comparative research between professionals and students.

AD	#Versions	#Words
English	10	6,814
Catalan	10	5,702
Spanish	10	5,292

Table 1: Number of words and audio descriptions.

5. Corpus annotation

After an analysis of various multimodal corpus analysis tools, ELAN[1] was selected to create complex annotation on the video resources (Sloetjes and Wittenburg, 2008). Essentially, ELAN allows linking annotations with their corresponding video files and saves these links in the annotation file. The annotation file is an XML file conforming to the EAF format[2]. ELAN also provides a powerful set of tools to assist video encoding and to perform eventual analyses, hence it was prioritized over other multimodal corpus analysis tools.

Corpus annotation, which is still ongoing, is designed at two main levels:

Linguistic annotations consist of an AD plus a set of dependent layers, where the AD is tied to the timeline and the dependent layers are tied to a specific annotation in the audio description itself. Six levels of linguistic dependent annotations have been included, namely: sentences, chunks, tokens, part of speech, lemma, and semantic annotations.

Sentence, chunk and token tiers are simply used to split the AD into smaller parts and, hence, their annotation value is a sub-string of the AD. Lemma, part of speech, and semantic annotation[3] are used to further annotate tokens. Linguistic annotations are automatically encoded using the Standford parser[4] (for English and Spanish) outside the ELAN tool and eventually added into the EAF file. To add Stanford annotations into the EAF files extensive use of the Pympi package[5] was made.

Cinematic annotations are currently being developed and are to be applied to the audiovisual content. They include 'text', 'sound' and 'camera' annotations.

Text annotations encode text on screen, be they the opening credits, subtitles or other text, both added at the postproduction stage or as part of the action (for instance, when a characters reads the contact list on a phone).

Sound annotations are particularly relevant for our research because they have a direct impact on where audio description can be included. They are used to identify silence, music, sound effects, and speech. Since sound annotations may overlap, four different tiers have been defined.

Camera annotations, currently being developed, will focus mainly on scene transitions, which often delimitate different spatial-temporal settings, and also on the cinematic technique of zooming, used to focus the audience attention towards an individual object.

The nature of our primary data, together with their corresponding annotation sets, give our eventual corpora a rather special character. As illustrated in Figure 1, the corpus contains a single short movie, in three different languages, which has been annotated according to 'filmic criteria' and a set of 30 different 'derived versions', each providing an AD. These 'derived versions' vary in language and provider and are further annotated in linguistic terms. In some way, our corpus constitutes a comparable corpus where up to 30 ADs are aligned against the same annotated timeline.

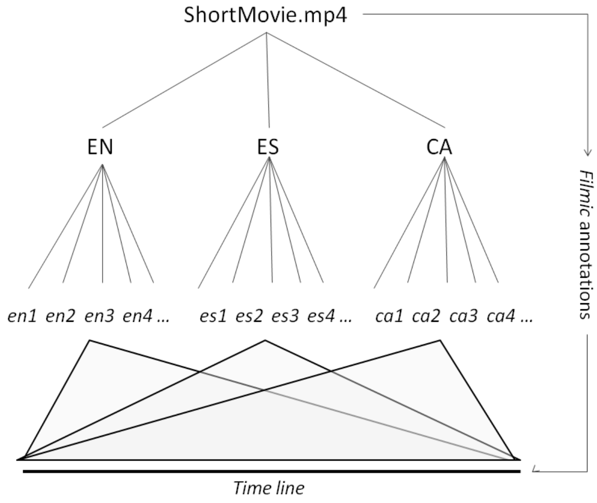

Figure 1: Corpus structure

6. Corpus exploitation

All these annotations allow a wide number of analyses to be performed. These may run on a particular file or on a set of files, permitting not only single analysis but also comparative analyses (for example, when comparing among languages or providers).

Figure 2 displays two different ADs (one from the United Kingdom and another from Canada) in the timeline. With this visualization, the researcher can easily see the annotations around a given point of time, quickly identify hot intervals and compare distributions between the two providers, among other features.

[1] See hppt://tla.mpi.nl/tools/tla-tools/elan, developed by the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

[2] The ELAN Annotation Format, also known as the EUDICO Annotation Format Netherlands.

[3] Currently used to classify verbs and adverbs.

[4] Stanford Lexicalized Parser v3.5.2 (http://nlp.stanford.edu/software/lex-parser.shtml).

[5] Pimpy is a package that allows manipulation of ELAN and TextGrid files.

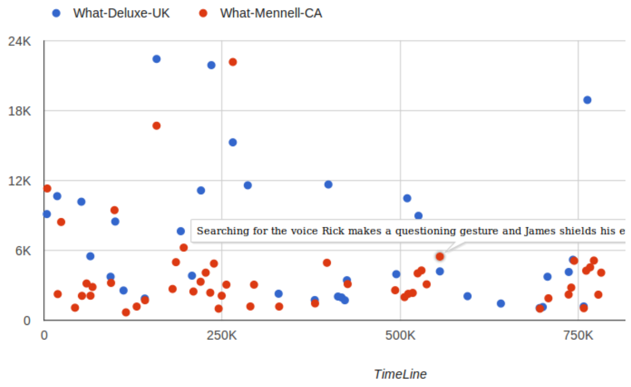

Figure 2: Sample visualization

Similarly, the set of annotations available enables a focus on a single layer or rather mixing different annotation layers. Thus, when considering linguistic layers alone, a variety of calculations can be performed on word frequencies, word distributions (both on timelines and among different EAF files), density, etc. Figure 3 shows the number of words/sentences per paragraph.

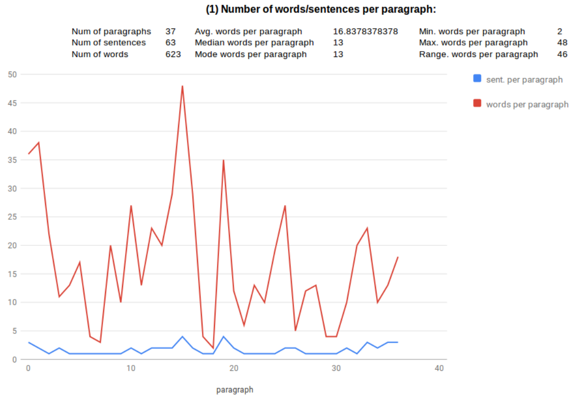

Figure 3: Word/sentence visualization

Particularly interesting is the correlation between camera and linguistic annotations. In this case, multilayer concordances allow identification of relevant correlations between significant filmic annotations such as shot transitions or zooms and the ADs. The fact that all annotations are aligned to the same annotated timeline opens up a wide range of possibilities.

The web application that is currently being developed provides access to source data and (some) graphic visualizations of that data. Source data include the ELAN annotation files as well as a set of csv files prepared to support the range of experiments and analyses that researchers may define. The graphic visualizations provided aim to explore and exemplify the possibilities of the annotated data available. In this case, Google Charts API is used to generate the charts out of the source data.

7. Conclusions

All in all, this paper has presented an ongoing project whose aim is to develop a corpus of AD created for a single film input. Despite its current limited size, our belief is that it is an innovative resource in terms of audiovisual text types and approach. It is multimodal and multilingual, and allows comparison at various levels of how the same visual input is translated into words in different languages and by different describers.

8. Acknowledgements

Project funded by Spanish Ministry within the Europa Excelencia programme (FFI2015-62522-ERC). Anna Matamala is a member of TransMedia Catalonia, funded by Catalan government funds (2014SGR027).

9. References

Baños, R., Bruti, S., Zanotti, S. (2013). Corpus linguistics and AVT: in search of an integrated approach. *Perspectives*, 21(4), pp. 483--490.

Bourne, J., Jiménez, C. (2007). From the visual to the verbal in two languages. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All*. Amsterdam: Rodopi, pp. 175--188.

Braun, S. (2013). Audiodescription research: state of the art and beyond. *Translation Studies in the New Millennium*, 6, pp. 14--30.

Chafe, W. (1980) (Ed.) *The Pear Stories*. Norwood: Ablex.

Jiménez Hurtado, C., Rodríguez, A., Seibel, C. (2010). *Un corpus de cine.* Granada: Tragacanto.

Jiménez Hurtado, C., Seibel, C. (2011). Multisemiotic and multimodal corpus analysis of audio description: TRACCE. In A. Remael, A., P. Orero, & M. Carroll (Eds.), *Audiovisual Translation and Media Accessibility at the Crossroads*. Amsterdam: Rodopi, pp. 409--425.

Mangiron, C., Maszerowska, A. (2014). Strategies for dealing with cultural references in AD. In A. Maszerowska, A. Matamala, & P. Orero (Eds.), *Audio Description*. Amsterdam: Benjamins, pp. 159--178.

Maszerowska, A., Matamala, A., Orero, P. (Eds.) (2015.). *Audio Description*. Amsterdam: Benjamins.

Matamala, A., Orero, P. (2013). Standardising audio description. *IJSEI*, 1, pp. 149--155.

Mazur, I., Kruger, J.-L. (2012). Pear Stories and Audio Description: Language, Perception and Cognition across Cultures. *Perspectives*, 20(1), pp. 1--3.

Orero, P. (2007). Sampling Audio Description in Europe. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All.*. Amsterdam: Rodopi, pp. 111--125.

Piety, P. (2004). The language system of audio description: an investigation as a discursive process. *JVIB*, 98(9), pp. 453-469.

Remael, A., Reviers, N., Vercauteren, G. (Eds.) (2015.). *Pictures painted in words: ADLAB Audio Description guidelines*. Trieste: EUT.

Salway, A. (2007). A corpus-based analysis of AD. In J. Díaz-Cintas, P. Orero, & A. Remael (Eds.), *Media for All*. Amsterdam: Rodopi, pp. 151--174.

Sloetjes, H., Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.