Purpose and features of the TVseriesAD corpus

Eva Schaeffer-Lacroix, Nathalie Mälzer, Kirsten Berland, Saskia Schulz

Eva Schaeffer-Lacroix, Nathalie Mälzer, Kirsten Berland & Saskia Schulz

Purpose and features of the TVseriesAD corpus

Abstract

This paper is related to the TADS (Translation of Audio Description Scripts) project, designed by a team of French and German researchers from Universität Hildesheim and Sorbonne Université. It focuses on the purpose and the features of the French-German TVseriesAD corpus. In our presentation, we outline the role this corpus is supposed to play for the development of a technology-assisted translation method for audio description scripts. With the help of one of our pilot datasets, the Buettenwarder subcorpus, we will demonstrate the way our complete data set will be tagged, and we will explain what sort of items can be searched in the corpus.

1 Introduction

Audio description (AD) scripts are texts composed of film script prompts, time indications, speaking or acting indications for the audio describer, and the text he or she has to record in order to create an AD audio track. Inspired by López Vera (2006), Jimenez Hurtado and Soler Gallego (2013), and Jankowska (2015), our project aims to identify procedures which support the semi-automatic translation of AD scripts. One of our research interests is to develop a method to distinguish between standardised features of French and German TV series AD scripts likely to be mastered by machine translation and other features needing human post-edition or direct human translation.

2 Research questions and hypotheses

Our research project relies on two assumptions:

1) The translation of AD manuscripts is a relevant option to increase the AD offer for the language pairs French-German, German-French.
2) There are AD script contents which are suited for machine translation and others which should better be translated by humans to maintain a high-quality standard.

This is our research question: To which extent can the AD script translation workflow be optimised with the help of machine translation without loss of quality?

Table 1 presents our hypotheses linked to this question and it lists the methods which will be applied to test them.

Hypothesis	**Applied method**
AD scripts, especially those of TV series, contain strongly standardised, recurrent elements (Kleege 2016).	Create an AD script corpus to identify and compare the linguistic characteristics of AD scripts of TV series with AD scripts of movies and short feature films.

Machine translation (MT) leads to good results for translating recurrent elements of AD scripts (Fernández-Torné 2016).	Use of a translation memory containing recurrent AD-specific chunks identified in the AD script corpus to translate a dubbed video in three different ways: HT with viewing the video, HT without viewing the video, MT.
Localisation tasks (cultural adaptation and user orientation) are better handled by human translators (HT) (Fernández-Torné and Matamala 2016).	Translate a sample of selected AD script parts containing localisation issues or elements referring to emotion. Apply three different translation methods: HT with viewing the video, HT without viewing the video, MT.

Table 1. Hypotheses and applied methods.

3 The TVseriesAD corpus

3.1 Content

In order to study the potential benefits of machine translation, Fernández-Torné and Matamala (2016) compared the efforts required to create, translate, and post-edit audio descriptions. Their conclusions open the field for several studies, amongst them the output quality of machine translations compared to the quality of human translations. Our project includes the study of lexical databases for audio describers, AD guidelines (e.g. those edited by Remael, Reviers, and Vercauteren n.d.), and audio description analysis methods (Fix 2005). These steps are followed by the corpus-based exploration of the linguistic features of AD scripts written by humans (cf. Salway 2007). Finally, our results will be compared to those obtained by the semi-automatic translation method to be developed.

At the present stage of our research, the corpus includes two German datasets, made available for teaching and research purposes by the regional broadcasting channels NDR (Norddeutscher Rundfunk) and BR (Bayerischer Rundfunk): (1) the Buettenwarder corpus, containing the audio description scripts of 69 episodes of the German TV series *Neues aus Büttenwarder* (Eberlein 1997) telling rural stories from Northern Germany, provided by Ursula Heerdegen-Wessel; (2) the Dahoam corpus, containing 605 episodes of the Bavarian family series *Dahoam is Dahoam* (Schmidt-Märkl *et al.* 2007) provided by Bernd Benecke. This starting corpus contains 1,888,393 tokens, punctuation marks included.

German original	English translation
10:00:38 s Am Dorfteich sitzen Adsche und Brakelmann auf einer Bank. 10:01:26 "Was ist das wieder herrlich ..." ("heute" übersprechen)	10:00:38 s Adsche and Brakelmann are sitting on a bench at the village pond. 10:01:26 "Isn't it lovely…?" (override "today")

Table 2. AD sample from the Buettenwarder corpus, episode 73.

The TVseriesAD corpus will be supplemented by French audio descriptions of TV series. In a second step, a more general corpus will be created, containing parallel and comparable German and French AD scripts for TV series, movies, and short feature films. To obtain relevant data, we started a collaboration with the broadcast companies Arte, France TV, NDR, KiKa, and SWR.

3.2 Data pre-processing

The pre-processing stage of AD scripts of television series involves significant challenges: they are composed of heterogeneous text elements, which are presented and distinguished from one another in various manners. In addition, the AD norms applied may vary throughout the files, in particular if their episodes are written by different authors, which is the case in the Buettenwarder data set.

In order to enable consistent and valid corpus queries, we did editorial interventions (e.g. correct typos) and we normalised a certain number of symbols (e.g. quotation marks). We detected varying AD conventions in the files corresponding to the episodes of the TV series. Based on the advice given by the German AD specialist Anke Nicolai, we made some modifications helping to reduce these differences: we deleted the scenery numbers; we inserted missing scenery change hash tags; we normalised the AD signs <s> and <ss> by deleting surrounding brackets; we replaced the plus signs surrounding dialogue prompts with double quotation marks; finally, we replaced the plus signs surrounding speaking instructions with round brackets.

3.3 Data annotation

We used the Buettenwarder subcorpus as a sample to develop our data annotation procedures. For analysing our German and French data and our planned parallel subcorpora (e.g. the audio description scripts of Sven Taddicken's movie Emmas Glück [Emma's Bliss] and the French translation of these scripts), we decided to use TXM (Heiden 2010), an interoperable open source platform for text analysis. During the upload of our Buettenwarder data to TXM, they were automatically annotated on the part-of-speech level with the help of the TreeTagger (Schmid 1994).

Item	**Tagging mode**	**Tool or technique**
parts of speech	automatic & post-editing	TreeTagger
AD-specific items	manual or semi-automatic	TreeTagger, XML

Table 3. Tagging choices.

In addition, to structure the Buettenwarder dataset, we enriched it with XML-TEI tags (Burnard 2014), in order to distinguish between film script prompts, time indications, speaking or acting indications for the audio describer, and the text to be recorded (see Table 4). Thanks to the provided structure, these different parts can be searched and described separately.

element	**XML-TEI tag**	**Buettenwarder sample (episode 12)**
time indications	<time>	10: 06: 50
parts to be recorded (*speaker*)	<sp>	Brakelmann fährt durchs Dorf. [Brakelmann drives through the

		village.]
instructions what to do or not during AD recording	<stage>	(Rest übersprechen) [Talk over the rest]
pace	<stage type="delivery">	s, ss, n [fast, very fast, normal]
displayed text (titles, signboards)	<caption>	Schild am Straßenrand. "Goethe Eier 5 Euro 10" [Sign at the roadside. "Goethe-eggs for 5.10 Euros]
(end of) dialogue lines	<prompt>	"Goethe gehört mir!" [Goethe is mine!]

Table 4. XML annotation sample.

4 Linguistic features

To identify the idiosyncratic linguistic features of the Buettenwarder subcorpus, we adopted a procedure based on results of Reviers et al. (2015: 2-3), who identified part-of-speech features which characterise Dutch AD film scripts. They conclude that "there is a 'language of audio description' that differs considerably from general language" (p. 17). Some of our findings are in line with those cited by Reviers et al. (2015: 5-6): the Buettenwarder corpus contains a high proportion of short sentences, and subordinates are rare. Salient features are verb particles, prepositions, compound adjectives and present participles. Verbs (except those used in the prompts) appear nearly exclusively in the third person of the present indicative tense, and some of the utterances do not contain any verb: "Jürgen in die Kamera: (...) [Jürgen towards the camera: (...)]".

We decided to concentrate not only on AD-specific parts-of-speech but also on n-grams and word sketches, informing on the words' collocational behaviour. This is motivated by our research focus, namely (besides evaluating AD script quality) identifying standard-like elements in the corpus (Kleege 2016) suited for machine translation. The n-gram function provides the corpus user with a list of frequent continuous word sequences.

n-gram	**abs. frequ.**	**n-gram**	**abs. frequ.**
ss sehr schnell sprechen	69	zieht die Augenbrauen hoch	15
ss speak very quickly	69	raises the eyebrows	15
durch Betonung deutlich machen	69	kommt auf dem Mofa	15
stress through intonation	69	arrives on his moped	15

Table 5. Absolute frequency of n-grams.

We searched the Buettenwarder corpus for n-grams (length: 4 to 6 words). The left column in Table 5 presents items corresponding to speaker instructions and the right column lists items to be recorded by the AD speaker.

5 Preliminary results

Normalising the Buettenwarder data was a long, but necessary and insightful pre-processing task. We had to take a close look to the text and sometimes even to compare a text part with the corresponding moment in the film. We were struck by the wide variety of AD conventions in the 69 episodes and we understood that it is a big challenge to establish national (or even international) formal standards for AD script creation. The post-editing effort for automatic annotation was bigger than expected, whereas the semi-automatic annotation process was relatively easy to cope with.

6 Further steps

On the basis of the results obtained, we plan to improve our procedures (how to annotate the data and to identify standardised features) before applying them to the whole AD corpus. These tasks will be followed by the analysis of non-standardised features, e.g. localisation issues and elements linked to emotion.

Acknowledgments

Our warmest thanks go to Ursula Heerdegen-Wessel from NDR (Norddeutscher Rundfunk) and Bernd Benecke from BR (Bayerischer Rundfunk) for having provided us with the datasets cited in this article.

References

Burnard, Lou. 2014. *What Is the Text Encoding Initiative? - OpenEdition Press*. http://books.openedition.org/oep/426.

Fernández-Torné, Anna. 2016. 'Machine Translation Evaluation through Post-Editing Measures in Audio Description'. *InTRAlinea* 18. http://www.intralinea.org/archive/article/2200.

Fernández-Torné, Anna, and Anna Matamala. 2016. 'Machine Translation in Audio Description? Comparing Creation, Translation and Post-Editing Efforts'. *SKASE Journal of Translation and Interpretation* 9 (1): 64–87.

Fix, Ulla, ed. 2005. *Hörfilm: Bildkompensation durch Sprache: Linguistisch-filmisch-semiotische Untersuchungen zur Leistung der Audiodeskription in Hörfilmen am Beispiel des Films 'Laura, Mein Engel' aus der 'Tatort'-Reihe.* Philologische Studien und Quellen, 189. Berlin: Erich Schmidt.

Heiden, Serge. 2010. 'The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme'. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 389–398. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://www.aclweb.org/anthology/Y10-1044.

Jankowska, Anna. 2015. *Translating Audio Description Scripts*. Bern: Peter Lang Edition. https://doi.org/10.3726/978-3-653-04534-5.

Jimenez Hurtado, Catalina, and Silvia Soler Gallego. 2013. 'Multimodality, Translation and Accessibility: A Corpus-Based Study of Audio Description'. *Perspectives* 21 (4): 577–94. https://doi.org/10.1080/0907676X.2013.831921.

Kleege, Georgina. 2016. 'Audio Description Described: Current Standards, Future Innovations, Larger Implications'. *Representations* 135 (1): 89–101. https://doi.org/10.1525/rep.2016.135.1.89.

López Vera, Juan Francisco. 2006. 'Translating Audio Description Scripts: The Way Forward? - Tentative First Stage Project Results'. In *Audiovisual Translation Scenarios: Conference Proceedings*, 10. Copenhagen. http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Lopez_Vera_Juan_Francisco.pdf.

Remael, Aline, Nina Reviers, and Gert Vercauteren, eds. n.d. 'Pictures Painted in Words - ADLAB Audio Description Guidelines'. Accessed 20 November 2019. http://www.adlabproject.eu/Docs/adlab%20book/index.html.

Reviers, Nina, Aline Remael, Walter Daelemans, Anna Jankowska, and Agnieszka Szarkowska. 2015. 'The Language of Audio Description in Dutch: Results of a Corpus Study'. In *New Points of View on Audiovisual Translation and Media Accessibility*, 167–90. Peter Lang. Avialable at http://www.clips.ua.ac.be/~walter/papers/2015/rrd15.pdf.

Salway, Andrew. 2007. 'A Corpus-Based Analysis of Audio Description'. In *Media for All: Subtitling for the Deaf, Audio Description, and Sign Language*, edited by Jorge Cintas Díaz, Pilar Orero, and Aline Remael, 151–74. Amsterdam: Rodopi.

Schmid, Helmut. 1994. 'Probabilistic Part-of-Speech Tagging Using Decision Trees.' In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

TV series

Schmidt-Märkl, Markus et al., directors. 2007. *Dahoam is Dahoam*. Bayerischer Rundfunk.
Eberlein, Norbert, director. 1997. *Neues aus Büttenwarder*. Norddeutscher Rundfunk.

Authors

Eva Schaeffer-Lacroix is a senior lecturer at the Department of Education of Sorbonne Université (Paris, France), where she teaches applied linguistics, ICT (Information and Communications Technology), and German as a foreign language. Her main research interests are corpus linguistics, writing in a foreign language, audio description, and translation.

Email: eva.lacroix@sorbonne-universite.fr
Website: http://didaktik.hautetfort.com/

Nathalie Mälzer is a professor for Transmedial Translation at Universität Hildesheim in Germany, where she developed a master program focusing on audiovisual translation: "Medientext und Medienübersetzung". Her research interests are audiovisual translation, accessibility (especially audio description and SDH), and literary translation. She translated more than 40 novels, plays and non-fiction books from French into German.

Email: maelzers@uni-hildesheim.de
Website: https://www.uni-hildesheim.de/fb3/institute/institut-fuer-uebersetzungswiss-fachkommunikation/mitglieder-des-instituts/maelzer/

Kirsten Berland holds a bachelor's degree in Korean Language and Multilingual Natural Language Processing at INALCO (Institut National des Langues et Civilisations Orientales), Paris (France). She is currently enrolled in the first year of the master program NLP (Natural Language Processing) at INALCO. From February to July 2020,

she participated in a TADS pre-project during an internship conducted by Eva Schaeffer-Lacroix.

Email: kirsten.berland@gmail.com

Saskia Schulz is a research assistant in the field of audiovisual translation at Universität Hildesheim. She obtained both her bachelor's and master's degree in Hildesheim and already taught two seminars and participated in several empirical studies on subtitling and theatre surtitling for the d/Deaf and hard of hearing during that time. Her main research interests are audio description, dubbing, subtitling, and surtitling.

Email: schul026@uni-hildesheim.de