Evaluating the effectiveness of machine translation of audio description: the results of two pilot studies in the English-Dutch language pair

Gert Vercauteren
Nina Reviers
Kim Steyaert

revista.tradumàtica
tecnologies de la traducció

Gert Vercauteren
University of Antwerp
gert.vercauteren@uantwerpen.be;
ORCID:
0000-0001-6711-2005

Nina Reviers
University of Antwerp
nina.reviers@uantwerpen.be;
ORCID:
0000-0003-0271-6662

Kim Steyaert
kim.steyaert@uantwerpen.be;
ORCID:
0000-0000-0000-0000

Abstract

The field of translation is undergoing various profound changes. On the one hand it is being thoroughly reshaped by the advent and constant improvement of new technologies. On the other hand, new forms of translation are starting to see the light of day in the wake of social and legal developments that require that products and content that are created, are accessible for everybody. One of these new forms of translation, is audio description (AD), a service that is aimed at making audiovisual content accessible to people with sight loss. New legislation requires that this content is accessible by 2025, which constitutes a tremendous task given the limited number of people that are at present trained as audio describers. A possible solution would be to use machine translation to translate existing audio descriptions into different languages. Since AD is characterized by short sentences and simple, concrete language, it could be a good candidate for machine translation. In the present study, we want to test this hypothesis for the English-Dutch language pair. Three 30 minute AD excerpts of different Dutch movies that were originally audio described in English, were translated into Dutch using DeepL. The translations were analysed using the harmonized DQF-MQM error typology and taking into account the specific multimodal nature of the source text and the intersemiotic dimension of the original audio description process. The analysis showed that the MT output had a relatively high error rate, particularly in the categories of Accuracy – mistranslation and Fluency – grammar. This seems to indicate that extensive post-editing will be needed, before the text can be used in a professional context.

Keywords: media accessibility; audio description; machine translation; translation

Resum

El camp de la traducció està experimentant canvis profunds. D'una banda, està patint una autèntica transformació gràcies a l'arribada i la millora constant de les noves tecnologies. De l'altra, noves formes de traducció comencen a veure la llum arran de l'evolució

Rebuda: 31 de gener de 2021 / Acceptació: 14 de setembre de 2021 / Publicació avançada: 15 de desembre de 2021

Revista Tradumàtica. Número 19 (desembre 2021).
Cita recomanada: VERCAUTEREN, GERT; REVIERS, NINA; STEYAERT, KIM; (2021). «Evaluating the effectiveness of machine translation of audio description: the results of two pilot studies in the English-Dutch language pair». Revista Tradumàtica. Tecnologies de la Traducció, 19, 226-252. <https://doi.org/10.5565/rev/tradumatica.288>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

social i legal que exigeix que els productes i continguts que es creen siguin accessibles per a tothom. Una d'aquestes noves formes de traducció és l'audiodescripció (AD), un servei que té com a objectiu fer accessibles els continguts audiovisuals a les persones amb pèrdua de visió. La nova legislació exigeix que aquests continguts siguin accessibles abans del 2025, la qual cosa constitueix una tasca immensa atès el nombre limitat de persones que actualment tenen formació com a audiodescriptors. Una possible solució seria utilitzar la traducció automàtica per traduir les audiodescripcions ja existents a diferents idiomes. L'AD, caracteritzada per frases curtes i un llenguatge senzill i concret, podria ser una bona candidata per a la traducció automàtica. En aquest estudi pretenem demostrar la hipòtesi per a la combinació lingüística anglès-holandès. Concretament, fragments de 30 minuts d'AD de tres pel·lícules holandeses que es van audiodescriure originalment en anglès, han estat traduïts a l'holandès per mitjà de l'eina DeepL. Les traduccions s'han analitzat utilitzant la tipologia d'error harmonitzada DQF-MQM i tenint en compte la naturalesa multimodal específica del text font i la dimensió intersemiòtica del procés d'audiodescripció original. L'anàlisi ha mostrat que la producció de TA té una taxa d'error relativament alta, especialment en les categories de precisió –errors de traducció– i fluïdesa gramatical. Això sembla indicar que caldrà una extensa postedició, abans que el text es pugui utilitzar en un context professional.

Paraules clau: accessibilitat als mitjans; audiodescripció; traducció automàtica; traducció

Resumen

El campo de la traducción está experimentando diversos cambios profundos. Por un lado, la llegada de nuevas tecnologías en constante desarrollo la está reconfigurando completamente. Por otro, están empezando a surgir nuevas formas de traducción como resultado de avances sociales y legales que establecen que los productos y contenidos que se crean han de ser accesibles para todo el mundo. Una de estas nuevas formas de traducción es la audiodescripción (AD), un servicio orientado a hacer accesible para las personas con pérdida de visión los contenidos audiovisuales. La nueva legislación establece que estos deberán ser accesibles para 2025, lo que representa una tarea desmesurada, dado el limitado número de personas actualmente cualificadas en AD. Una posible solución pasaría por utilizar la traducción automática para traducir las audiodescripciones existentes a diferentes idiomas. Teniendo en cuenta que la AD se caracteriza por el uso de frases cortas y un lenguaje sencillo y concreto, sería una buena candidata para la traducción automática. El presente estudio ha puesto a prueba esta hipótesis con la combinación lingüística inglés-holandés. Para ello, se utilizó DeepL para traducir al holandés tres fragmentos de 30 minutos de AD extraidos de varias películas holandesas que habían

sido audiodescritas en inglés. Para analizar las traducciones se utilizó la tipología de error DQF-MQM armonizada y se tuvo en cuenta la naturaleza multimodal específica del texto de origen y la dimensión intersemiótica del proceso original de audiodescripción. El análisis mostró que el resultado de la TA presentaba una tasa de error relativamente elevada, especialmente en las categorías de Exactitud/Error de traducción y Fluidez/Gramática. Esto parece indicar que, antes de que el texto pueda ser utilizado en un contexto profesional, será necesario un proceso de posedición exhaustive.

Palabras clave: accesibilidad a los medios; audiodescripción; traducción automática; traducción

1. Introduction

Language technologies have had a profound impact on the field of Translation Studies. Globalization and digitization have made society at large ever more aware of the role of technology in the translation process, particularly in (digital) media and audio-visual products. The introduction of machine translation systems has been one of the major driving forces in this development. Since the turn of the millennium the advent of machine translation (MT) has significantly changed the way in which we translate (Bywood et al., 2017; O'Hagan, 2019; 2020). Over the last few years, concerns about MT as a threat to the translator's profession have given way to a more appropriate recognition of the active mediating role this technology takes in the translation process (O'Hagan, 2020). Indeed, the question is no longer whether or not we will accept MT as an alternative to translation from scratch, but how we can integrate it into our workflows and how it can improve both the quality and efficiency of the translation process (O'Hagan, 2019; 2020).

Given the growing body of legislation that makes it mandatory for broadcasters and other providers to make their audio-visual productions accessible, the question of the usability of MT is gaining relevance in the field of media accessibility as well. A key factor in the discussion for (commercial) market players is the challenge of balancing speed, quality and cost. Media service providers have to increase the access of their content through, for instance, audio description (AD) to offer people with a visual impairment equitable access to information and entertainment stipulated by a growing body of European and local legislation. In March 2019, for instance, the European Union adopted a new Accessibility Act (EAA)1 and an update of the European Audiovisual Media Services Directive (AVMSD).2 The EAA requires companies to make their websites, software and apps accessible within five years from the adoption of the act, which includes accessibility for people with sight loss, while the AVMDS states that member states have to make media services accessible to people with sight or hearing loss. Although these

1 Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services

2 Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU concerning the provision of audio-visual media services (Audiovisual Media Services Directive).

directives do not directly mandate the provision of AD and only time will tell how each member state translates the directives into stringent national legislation, it is clear that over the next few years public and private bodies will have to drastically step up their efforts to make audio-visual content in Europe more accessible for people with sight loss, by providing services such as AD. This constitutes a major challenge, particularly for smaller languages such as Dutch, for which AD is still a relatively new phenomenon (Reviers, 2016).

One way to meet these new quantitative demands is by translating existing ADs. Fernández-Torné and Matamala (2016) observe that ADs are generally still created as intersemiotic translations of the original audio-visual text, and interlingual translations of existing descriptions are very rare exceptions to that rule. This may seem surprising since there are still more (audio-visual) translators trained in interlingual translation than in AD, which would make translation a logical choice over description to rapidly increase the amount of audio described audio-visual content. In addition, translating ADs not only offers potential for meeting new market demands quickly. It is also a feasible solution for private broadcasters and companies in regions like Flanders and the Netherlands who distribute a lot of non-Dutch content with subtitles but no Dutch AD. In such cases, people with sight loss largely remain excluded from social and cultural life since they can only access the audio subtitles provided, i.e., a spoken version of the on-screen subtitles. Translations of non-Dutch ADs that can be used in combination with audio subtitles is therefore a crucial consideration.

In other words, it would be interesting to explore the idea of translating ADs in combination with the use of machine translation systems. However, very little scientific research and systematic evaluation are currently available to support the application of MT for AD. While there have been several studies on machine translation for subtitling (Matusov et al., 2019; Matamala and Ortiz-Boix, 2016; Álvarez Muniain et al., 2016; Etchegoyhen et. al., 2014; Del Pozo, 2013) and several machine translation subtitling systems are slowly finding application in the industry, machine translation systems for media accessibility and AD in particular have rarely been developed and have not been studied as yet. Only a handful of publications so far report on research into the translation of ADs (see section 2) and, to the best of our knowledge, only two small-scale exploratory studies focusing on the Spanish-Catalan and English-Catalan language pairs, have published results on MT for AD (Fernandez-Torné and Matamala, 2016; Matamala and Ortiz-Boix, 2016).

Against this background, this paper addresses the use of machine translation for AD. It reports on a case study conducted in 2019 by two students in the Master in Translation of the University of Antwerp (Uiterwijk, 2019; Bryssinck, 2019) and replicated by the authors of this study in 2020. In both the initial case study and the replicated study, the machine translations of three excerpts from English AD scripts into Dutch were manually evaluated. For this study, the neural MT solution of DeepL[3] was used. The focus of both case studies was on identifying the types of errors that occur in the Dutch

[3] https://www.deepl.com/translator

MT and evaluating the extent to which the most common types of errors correlate with or can be explained by the idiosyncrasies of AD as a text type.

This paper starts with a discussion of the current state of the art in (machine) translation of AD. Then, we will present the analytical framework for the case studies and the methodology adopted for error categorization. Finally, we will present a thorough discussion of the results and correlate them to the specific characteristics of the AD text type.

2. State of the art

AD translation, including MT of AD, is an area that has not received a lot of academic attention so far, given its still limited application in practice. Only a handful of publications address research into the *human* translation of ADs. The focus of both Jankowska (2015) and Lopez Vera's (2006) studies was on evaluating the efficiency and effectiveness of human AD translation as compared to creating an AD from scratch. Both studies reported positively about AD translation as an alternative to AD creation: it is less time-consuming and can help to increase the offer of AD quickly and cost-efficiently. More importantly, AD translation seems to generate good quality results. Jankowska (2015) and Herrador Molina (2006) tested the reception of translations of ADs with users in Poland and Spain respectively. In both studies the translations from English were evaluated positively overall. As a result, Jankowska concludes that "the scripts created as a result of the strategy of translating can be at least equal in quality to those which are the results of the strategy of writing." (Jankowska, 2015, p. 117).

A series of other studies (Herrador Molina, 2006; Bourne and Hurtado, 2007; Jankowska, 2015; Remael and Vercauteren, 2010; Jankowska, et. al., 2017; Liu, Tor-Carroggio, Rovira-Esteva and Casas-Tost, 2021) have looked into the actual translation process, discussing a series of AD-specific translation problems in different language pairs, namely from Polish into English and from English into Spanish, Dutch and Chinese. These preliminary studies have flagged a few potential translation crisis points (Pedersen, 2008, p. 101) – i.e., problematic passages that require active decision-making on the part of the translator such as linguistic, syntactical and cultural differences. While it is still too early to develop any general frameworks based on the above studies, they do point to potential AD-specific translation problems that are relevant for both the study of human translation and machine translation. These are issues AD translators should be aware of when they adapt their translation to the respective target audience. In the case of MT, these issues would have to be checked during the post-editing process.

First, there are linguistic and stylistic differences in the way AD is formulated between language pairs. An example is sentence length and complexity. Bourne and Hurtado (2007), Molina (2006) and Remael and Vercauteren (2010) all mention that complex English sentences from the source text were often adapted in the human translated TT into coordinating sentences or a series of simple sentences. Remael and Vercauteren (2010) also noticed the frequent use of the present participle in English AD, a grammatical form that cannot be easily transferred to Dutch where it is used in a different way. To

give one final example, Liu, Tor-Carroggio, Rovira-Esteva and Casas-Tost (2020) noticed differences between English and Chinese scripts in terms of the level of explicitness, the way characters are named and described and how much information is conveyed in the AD.

Second, the audio described text that is translated is part of a larger multimodal text with which it interacts on many levels (Reviers, 2018b). Indeed, the audio-visual product that constitutes the source text of the AD is a multimodal construct that, in addition to images and sound effects, generally contains dialogues in the source language. As such it seems logical that ADs are directly created in the same language as these dialogues rather than in a different one – by people who are not native speakers of that original language – in order to guarantee maximum inter- and multimodal coherence between the sound effects, dialogues and descriptions. The above studies mention that this might be an issue for AD translation. One example are cultural references. As Remael and Vercauteren (2010) point out, particularly in the case of AD, at least part of the translation problem of cultural references is closely related to the visual context/source. The way a cultural reference can be translated depends on the way in which it is simultaneously depicted on screen or audible in the soundtrack or in the audio subtitles (AST) or dubbed dialogue. However, the multimodal nature of the source text is likely also to impact on other levels of AD translation than cultural references alone. Previous research into the language of AD (such as Reviers, 2018a) shows the close interaction between an AD and the sounds and dialogues with which it is combined. This suggests that when translating ADs, this multimodal cohesion is a key feature to keep in mind as it needs to remain intact in the translated version as well.

A final issue is timing. The number of words used in the source text and its translation may differ from one language to another: for example, a Dutch translation of an AD may be a few words longer than the original English version. While this may not constitute a problem in more traditional instances of interlingual translation, it is a crucial element in the multi- and intersemiotic context of AD translation, since ADs – like other forms of audio-visual translation such as subtitling and dubbing – always have to be adapted to the time available between dialogues and sound effects. Depending on the language pair, this could mean that the translation of the existing description has to be shortened and that, in some cases, information may have to be omitted.

To the best of our knowledge, two studies on MT of AD have been published so far (Fernández-Torné and Matamala, 2016; Matamala and Ortiz-Boix, 2016). A third study on the MT translation of AD between the French-German language pair is currently being conducted in the TADS project (Mälzer and Schaeffer, 2021), but results have not been published at the time of writing. Fernández-Torné and Matamala (2016) conducted a study about the implementation of MT in AD for the English-Catalan language pair, focusing on the process of post-editing versus translation from scratch. The authors compared three scenarios: AD creation in Catalan, AD translation from English into Catalan and AD MT post-editing from English into Catalan. They found that post-editing did not necessarily save a significant amount of time compared to human translation of the existing AD, but other objective measures such as technical effort as measured

through keyboard and mouse interaction, and cognitive effort as measured through pause to word ratio, "seem to be less demanding in post editing" (Fernández-Torné and Matamala, 2016, p. 80). On the other hand, subjective indicators suggested that post-editing was perceived to be the most demanding task when compared to creating AD from scratch and translating existing ADs. In the interviews conducted after the experiment, participants indicated that they felt their creativity was impaired when they had to perform post-editing of MT output.

Matamala and Ortiz-Boix (2016) subsequently conducted a study focusing on the effectiveness of machine translation for the translation of AD scripts for the Catalan-Spanish language pair. Their corpus consisted of the Catalan AD from the first episode of a Catalan television series and the Catalan AD of a movie, resulting in approximately 90 minutes and 4,384 words of AD (Matamala and Ortiz-Boix, 2016, p. 17). They opted to carry out a subjective evaluation performed by a human, based on a list of error types which would be looked at (Matamala and Ortiz-Boix, 2016). These errors comprised missing words, untranslated words, extra words, wrong word order, wrong agreement, incorrect words and mistranslated words. The errors, in descending order of frequency, were wrong word order, wrong agreement, incorrect words, mistranslated words, untranslated words and missing words. They also reported that about half of the output sentences contained at least one error (Matamala and Ortiz-Boix, 2016).

To conclude, the previous studies are inconclusive when it comes to both the efficiency and the effectiveness of MT for AD. Matamala and Ortiz-Boix (2016) suggest that the post-editing effort required to bring MT of AD to an acceptable level of quality is considerable, given the high number of errors. More research is required into both the post-editing process as well as the types of errors most frequently encountered and the reasons for these errors.

3. Methodology

Since the main aim of the present study is to obtain preliminary insight into the types of mistakes that can be found in machine translations of English ADs into Dutch, various specific parameters in terms of materials and assessment procedure were taken into account when designing the methodology. The aspects discussed below apply to both the initial case study of 2019 by Uiterwijck and Bryssink as well as its replication by the authors of the present paper in 2020.

3.1. Materials

The existing corpus of Dutch ADs that are translations of ADs originally created in another language is still very limited. As mentioned in the introduction, ADs are mainly created from scratch in the same language as the audio-visual product. In the case of Dutch, that means that Dutch ADs are created only for Dutch productions. ADs from the UK or the US are rarely translated into Dutch, even though the amount of content with AD from these countries far exceeds the numbers of ADs created in Flanders and the Netherlands combined (Reviers, 2016).

For the purpose of our analysis, three Dutch feature films that were originally described in English were selected, namely *Blind* (Van den Dop, 2007), *Zwartboek* (Verhoeven, 2006) and *Het leven is vurrukkulluk* (Weisz, 2018). At the time of the study, these were virtually the only translated ADs available for this language pair. For each of the three films, the English AD of the first 30 minutes was transcribed and segmented into individual AD blocks, yielding 131 AD blocks or 2,600 words for *Blind* (Van den Dop, 2007), 120 blocks or 2,051 words for *Zwartboek* (Verhoeven, 2006) and 86 blocks or 1,400 words for *Het leven is vurrukkulluk* (Weisz, 2018). Since a pre-analysis of the MT of the AD of *Zwartboek* (Verhoeven, 2006) indicated that there were more than 150 errors in the first 30 minutes of the AD, it was decided to limit the case-study to 3 times 30 minutes to keep the workload involved in the human quality assessment feasible for this first pilot study. In addition, it resulted in a data set that was comparable to that of the earlier case study by Matamala and Ortiz-Boix (2016) in terms of AD time and word count.

3.2. MT Engine

For the machine translation of the source data, it was decided to use the general and freely available Neural MT engine DeepL for both case studies. Several reasons guided this choice. First, we wanted a system that was freely accessible online for reasons of replicability and availability. The main freeware engines available are DeepL and Google Translate. A pre-analysis of the MT of one of the three films selected did not generate significant differences in terms of the number of errors between DeepL and Google Translate for the text selected. Given that it was beyond the scope of our study to perform an in-depth comparison between the two engines, we opted for DeepL, but for the future it may be worthwhile to compare Google Translate and DeepL, for example in terms of types of errors made and post-editing effort required, to see if one is more suitable for MT of AD than the other.

Second, we opted for a Neural Machine Translation system (NMT), because these are quickly becoming the standard in the industry over rule-based (RBMT) or statistical MT (SMT) systems. In NMT systems, one large neural network (NN) is trained on a vast amount of data consisting of full sentences and their translations. In these systems, the encoder maps the source sentence into a vector representation, and the decoder predicts the target sentence using that representation. In order to do this, the encoder-decoder network is jointly trained to maximize the probability of a correct target sentence, given a source sentence (Cho et al., 2014). This is different for SMT systems, which consist of many small sub-components that are tuned separately (Bahdanau et al., 2015). These systems use data to train a probabilistic model and choose the translation with the highest probability, given a certain source phrase. While these two systems heavily depend on large corpora, RBMT systems use extensive dictionaries and linguistic rules to translate sentences.

Third, at the time of writing, an MT system that was specifically trained for the translation of AD was not yet available. Furthermore, research on the linguistic and stylistic specificity of AD and its differences and similarities with other text types (such as novels, subtitling, spoken and written language) is too scarce to be able to arrive at

any solid hypotheses about what type of content would be most similar to AD (Reviers, 2018a; Arma, 2011) and, therefore, what would be most appropriate content to use as MT input. Some scholars have put forward the hypothesis that, due to the characteristics of the language of AD, general engines might be a good candidate for automatic translation (Fernández-Torné, 2016; Salway, 2004; 2007). Based on an analysis of existing guidelines, for instance, Vercauteren (2007) concludes that the language of AD "should sound natural and that unusual vocabulary or formal phrasing have to be avoided [...], sentences should be kept simple [...] and complex sentence structures including many subordinate clauses must be avoided" (p. 144). Similarly, Salway (2004) points out that "the relatively simple nature of the language used in audio description [...] may mean automatic translation systems fair (sic) better than usual." (p. 6). Reviers (2018a) came to a similar conclusion, based on the analysis of a corpus of Dutch ADs that also demonstrates the same features, such as simple sentences and concrete, non-specialised vocabulary. Fernández-Torné (2016) found that the SMT system of Google Translate outperformed other MT engines when translating English ADs into Catalan. However, such preliminary hypotheses have been drawn based on earlier MT technologies and have not yet been empirically tested so far, underlining the need for research in this area.

3.3. Translation Quality Assessment

As demonstrated by Castilho, Doherty, Gaspari, and Moorkens (2018), "MT quality can be assessed in a wide range of different ways, and no single approach or metric is sufficient to address all evaluation purposes and scenarios" (p. 24). Automatic systems, such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) or chrF3 (Popović, 2015)[4] are regularly used in various fields, ranging from technical texts to audio-visual content and literary MT to assess large scale productions and to analyse translation quality when the time that can be devoted to assessment is limited. While automatic quality assessment is said to be objective and tends to be inexpensive, it has also been claimed "that it is less comprehensive than manual evaluation and does not readily indicate the type of problems that the translated text contains" (Castilho, Doherty et al., 2018, p. 25). In light of this latter observation and since manual evaluation offers the means to obtain a fine-grained overview of the error types encountered in the translation (e.g., Popović, 2018; Lommel, 2018), which is relevant for the present study, we decided to evaluate the raw MT output generated by DeepL manually.

Earlier studies have resorted to both amateur evaluators and professionally trained evaluators. Castilho et al. (2018) point out that while "professional evaluators can be assumed to provide more reliable results, amateurs may be equally helpful in some TQA [Translation Quality Assessment] tasks" (p. 23). Moreover, Lommel (2018) indicates that an analytic evaluation – such as the one we are performing in this study – "is time-consuming and requires training for evaluators to apply consistently" (p. 122). Since there was no time to train evaluators for this study and since the main aim of the present study was to obtain a general overview of the main error types rather than an exact account of the number of errors in each category, the evaluation for our case study

4 Since the present article reports on a pilot study in which manual assessment was carried out, these automatic systems will not be discussed in more detail.

was carried out by the research team. The analysis was first undertaken by an MA student in Translation at the University of Antwerp and then replicated by the authors of the present study: it was checked by a PhD student in Translation Studies, specialising in AD, after which it was validated by two researchers specializing in AD.

3.4. Error analysis method

As discussed above, the English AD of the first 30 minutes of three films was transcribed and segmented into individual AD blocks, adding up to a total of 343 AD blocks (See table 1 below). The length of these AD blocks in the source text (ST) varied between two and ninety words, with an average of 17.5 words per AD block. The segmentation was based on the timecodes provided with the written AD scripts that were used to record the AD. A block of AD can consist of one word or several sentences and usually ends when (a) a character in the audio-visual production begins to speak, (b) there is a significant sound that cannot be covered by AD, or (c) there is a significant pause after a sequence of AD. Each of these AD blocks was then translated using DeepL by copying and pasting several blocks of the scripts at a time, up to the maximum allowed in the free version of the online engine.

Given the novelty of (machine) translation of AD, there are no frameworks as yet that are commonly used to evaluate the quality of the translated output. For their analysis of the translation of AD from Catalan into Spanish, Matamala and Ortiz-Boix (2016) created their own typology, as described in section 2. However, this typology is limited and does not fully take into account the specific multimodal context in which the translation of AD takes place. Therefore, for the translation quality assessment and error classification in our study, the harmonized DQF-MQM error typology[5] (DQF-MQM) was used. This error typology is the result of the integration of two similar frameworks, i.e. the TAUS Dynamic Quality Framework (DQF) and the Multidimensional Quality Metrics (MQM) framework, developed by DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH), into one hierarchic typology containing eight error categories and a total of 50 issue types (Lommel, 2018). For this study, two subcategories were added to category 8 *(Other)* of the DQF-MQM typology in order to accommodate the nature of the source text: *8.1 Mistake in source;* and *8.2 Unnecessary translation*. This comprehensive harmonized error typology, that was designed to evaluate both human and machine translations, allows for practical and easy error categorization and is one of the standards for quality assessment in both research and industry. For these reasons and with a view to replication in the future, this framework was preferred over the creation of an error typology from scratch or using the one developed by Matamala and Ortiz-Boix (2016).

The analysis was conducted in an Excel spreadsheet for each film. The first columns contain the two parallel texts, i.e. the English AD script divided in AD blocks aligned with the Dutch NMT, followed by columns containing the errors and the error annotation within the DQF-MQM framework.

5 https://www.taus.net/qt21-project#harmonized-error-typology

4. Analysis and discussion

In this section, we will discuss the results of the translation of the three AD scripts by DeepL and give an overview of the most common types of errors found in the output. With 520 marked errors on a target text (TT) of 6,374 words and 69.7% of all translated AD blocks containing errors (see table 1), it can be said that NMT does not deliver an output that is ready for use without thorough revision and post-editing. On a sentence level, this error rate is considerably higher than the error rates of the rule-based MT (RBMT) and statistical MT (SMT) used in the study by Matamala and Ortiz-Boix (2016), who found that 57.8 and 42.2% respectively of the machine translated sentences in their study contained at least one mistake. This marked difference can be explained by two factors. Firstly, we did not work with individual sentences but rather with AD blocks that often consisted of more than one sentence. Secondly, Matamala and Ortiz-Boix' experiment consisted of a language pair with two more closely related languages (Catalan-Spanish) than our language pair (English-Dutch). A new analysis would have to be conducted to assess how our results compare with those of Matamala and Ortiz-Boix (2016) on an individual sentence level. On a word level, however, our error rate of 8.16% with 520 marked errors among 6,374 translated words is in line with the findings of Matamala and Ortiz-Boix (ibid.) who found 11% for the RBMT and 5.56% for the SMT engine, averaging an error rate of 8%.

	Word count English AD	Word count Dutch NMT translation	AD blocks	AD blocks without errors	Total marked errors
Blind	2,616	2,819	134	37	252
Het leven is vurrukkulluk	1,400	1,445	88	31	102
Zwartboek	2,058	2,110	121	36	166
Total	6,074	6,374	343	104	520

Table 1 - Word count, AD blocks and number of errors

The distribution of errors, as illustrated in table 2, clearly shows that the main types of errors in the MT are in Accuracy or Fluency, with 86.1% of all errors falling in either of these categories. Less frequent errors are in Style and Other. Finally, since only 3 of 520 classified errors were included in the category Design (i.e., issues related to the design of the text, such as text length or formatting), these will not be discussed further.

	Accuracy	Fluency	Term	Style	Design	Locale	Verity	Other	Total
Blind	167	65	–	16	–	–	–	4	252
Het leven is vurrukkulluk	69	11	–	8	–	–	–	14	102
Zwartboek	116	21	–	20	3	–	–	6	166
Total	**352**	**97**	–	**44**	**3**	–	–	**24**	**520**

Table 2 - Distribution of errors per error type

4.1. Accuracy

The error category Accuracy combines all errors where the target text does not accurately reflect the source text (TAUS). Accuracy consists of seven subtypes. Most of the errors in our corpus were of this type and could be categorized under six of the seven subtypes, i.e. *addition, omission, mistranslation, over-translation, under-translation* and *untranslated.* From this list, *mistranslation* accounted for 85.1% of all accuracy errors, and therefore constitutes the most common error subtype (see table 3). No instances of the seventh subtype, *improper exact TM match,* occurred in this study since we did not use any translation memory (TM) to translate the AD scripts. The errors identified relate to misinterpreting the multiple meanings of homonyms, mistranslating sentences that contain a present participle or the conjunction 'as' to express simultaneity, and verb tense misinterpretation. This type of error can be attributed to not taking the context into account adequately.

	Addition	Omission	Mis-translation	Over-translation	Under-translation	Untranslated	Total
Blind	10	9	138	–	4	6	167
Het leven is vurrukkulluk	5	2	53	1	6	2	69
Zwartboek	8	5	96	–	3	4	116
Total	23	16	287	1	13	12	352

Table 3 - Distribution of accuracy errors per error subtype

In the following paragraphs we will present examples for the different error types mentioned. Each example consists of an example number, the title of the film, the AD

block, the original English sentence, the Dutch MT, and the English back translation of the Dutch MT or a post-edited version of the Dutch MT, depending on the type of error. It should be made clear that all the examples below have been chosen for reasons of clarity and in order to demonstrate a specific error type. In the event that several errors (of different types) occur in one example, only the error related to the error type under discussion will be analysed. This means some of the examples below contain more errors than only the one discussed as an example. This also means that in some of the blocks in the examples below, the error discussed is the only error present, which may lead one to believe the number of errors is generally rather low. On average, however, the AD blocks contain 2.3 errors, and only 30.3% of the AD blocks in this case study contain no errors at all. In other words, a considerable number of AD blocks contain more than one error, which is illustrated in example 18 in which five errors occur: two spelling errors (apostrophe s in "Rachel's" and "vader's"), an omission ("concealed"), an unidiomatic translation ("gouden staven" for "gold ingots" instead of the idiomatic "goudstaven"), and the awkward construction that is discussed in the example itself.

As mentioned above, the MT often misinterprets the multiple meanings of homonyms and returns a correct translation of the word itself, but incorrect in the immediate context of the word, as in the example below: The word "corporal" can refer both to the military function of a person as well as an adjective to refer to the human body. In the context of the film *Zwartboek*, corporal refers to a character, while it was erroneously translated into Dutch as "lichamelijk", referring to the human body. The verb "sniffs" in the same sentence was incorrectly translated as well by the Dutch noun "snuffels", referring to the sound that is made by the nose while sniffing.

1	Zwartboek	87	EN (original	The corporal sniffs
			NL (MT)	De lichamelijke snuffels
			EN (back translation)	The physical snuffles

In some cases, the mistranslation is not as apparent as in the previous example, as illustrated in example 2 below. Typical for AD is the multimodality of the medium (see section 2). Certain errors were not immediately identified as a mistranslation by the evaluators, but only later on while consulting the multimodal context. In the example below, "jumps" in the original text refers to a sudden movement caused by surprise. The MT, however, contains a Dutch word referring to one of the verb's other meanings, i.e., to jump in the air. In addition to this mistranslation, the noun "pace" has been omitted in the MT.

2	Blind	1-11	EN (original)	She slows her pace and jumps as crockery flies past her
			NL (MT)	Ze vertraagt haar en springt als het serviesgoed langs haar heen vliegt...
			EN (back translation)	She slows her and leaps as crockery flies past her...

Another recurrent error in the MT is the misinterpretation of the verb tense. In our analysis we found misinterpretations of both the verb tense (in example 3 the simple present is translated into a simple past) and the verb aspect (in example 4 the English past participle is translated into a Dutch infinitive).

3	Zwartboek	113	EN (original)	Suddenly a truck and German troops burst through the door. (simple present)
			NL (MT)	Plotseling barstten een vrachtwagen en Duitse troepen door de deur.
			EN (back translation)	Suddenly a truck and German troops burst through the door. (simple past)
4	Het leven is vurrukkulluk	82	EN (original)	Boelie steps into a bedroom and sees his mother passed out on the bed, an empty bottle of wine beside her.
			NL (MT)	Boelie stapt een slaapkamer binnen en ziet zijn moeder flauwvallen op het bed, een lege fles wijn naast haar.
			EN (back translation)	Boelie steps into a bedroom and sees his mother passing out on the bed, an empty bottle of wine beside her.

Distinctive for the language of English AD is the use of the present participle to describe simultaneous actions (Salway, 2007; Remael and Vercauteren, 2010). While the use of the present participle is also common in Dutch AD, it is rather uncommon in general texts or speech (Reviers, 2018a). As a result, DeepL translated the English present participle in different ways, often leading to mistranslations. The examples below show where DeepL mistranslated the simultaneity with, for instance, a wrong nominalization (example 5), or using a different word category (in example 6, the present participle is translated as a preposition).

5	Blind	1-39	EN (original)	Marie runs her fingers gently and slowly over the open book on the desk, tracing patterns with the tips of her fingers.
			NL (MT)	Marie loopt met haar vingers zachtjes en langzaam over het open boek op de bureau, het traceren van patronen met de uiteinden van haar vingers.
			EN (back translation)	Marie runs her fingers gently and slowly over the open book on the desk, the tracing of patterns with the tips of her fingers.
6	Blind	1-25	EN (original)	Passing a mirror she sees herself and averts her eyes.
			NL (MT)	Voorbij een spiegel ziet ze zichzelf en vermijdt haar ogen.
			EN (back translation)	Beyond a mirror she sees herself and averts her eyes.

However, DeepL also translated several sentences with the present participle correctly, finding an equivalent construction in Dutch that does not contain the present participle. In example 7, the verb has been replaced by a locative adverb, resulting in a natural sounding Dutch phrase.

7	Het leven is vurrukkulluk	51	EN (original)	She passes a bookcase and sees a woman of about 70, Rosa Overbeek, sitting at a desk reading a book with the aid of a magnifying glass.
			NL (MT)	Ze passeert een boekenkast en ziet een vrouw van ongeveer 70 jaar, Rosa Overbeek, achter een bureau een boek lezen met behulp van een vergrootglas.
			EN (back translation)	She passes a bookcase and sees a woman of about 70, Rosa Overbeek, behind a desk reading a book with the aid of a magnifying glass.

Not only present participles lead to translation errors. In their analysis of the human translation from English into Dutch from the films *Blind* (Van den Dop, 2007) and *Zwartboek* (Verhoeven, 2006), Vercauteren and Remael (2010) found that constructions in the English AD containing 'as' to express simultaneity were often translated incorrectly, too. This seems to be no different in machine translation. Most frequently, DeepL

translated 'as' by 'als', which entails a shift in meaning from simultaneity to conditional (example 8). However, in some AD blocks 'as' was translated correctly by 'terwijl', which according to Reviers (2018a) occurs with a high frequency in Dutch AD (example 9).

8	Zwartboek	24	EN (original)	As the bomber flies overhead it drops another bomb
			NL (MT)	Als de bommenwerper overvliegt laat hij een andere bom vallen.
			EN (back translation)	If the bomber flies overhead, it drops another bomb.
9	Het leven is vurrukkulluk	77	EN (original)	As they wait for Boelie, Mees leans in to give Panda a kiss, but she seems distracted by something.
			NL (MT)	Terwijl ze wachten op Boelie, leunt ze voorover om Panda een kus te geven, maar ze lijkt afgeleid door iets.
			EN (back translation)	While they are waiting for Boelie, she leans in to give Panda a kiss, but she seems distracted by something.

A final, less frequent, but, nevertheless, remarkable type of mistranslation is the creation of non-existent words by DeepL. This can be attributed to the characteristics of NMT, as NMT systems can operate on the level of subword units, as opposed to word-level MT models (Macken et al., 2020).

10	Blind	1-24	EN (original)	Catherine eyes her through a lorgnette then nods and stands aside to let Marie in.
			NL (MT)	Catherine kijkt haar door een vrachtwagengnet en knikt dan en gaat opzij staan om laat Marie binnen.
			EN (back translation)	Catherine looks her through a lorry net then nods and stands aside to let Marie in.

4.2. Fluency

All errors referring to issues related to the form or content of a text that are not directly related to the accuracy of its translation were classified under the category Fluency (DQF-MQM). As shown in table 4, most annotated fluency errors belong to the subtype

grammar (79.4%), as compared to spelling (8.2%) and punctuation (12.4%). Therefore, we will limit the examples to those with grammatical errors, such as word order errors, errors of subject-verb agreement, missing/incorrect preposition, wrong articles, and incorrect possessive case of names, which all undermine the fluency of the target text. In this section, each example consists of an example number, the title of the movie, the AD block, the original English sentence, the Dutch MT, and the corrected version of the Dutch MT, since the English back translation would not clarify the fluency error.

	Punctuation	Spelling	Grammar	Total
Blind	9	3	53	65
Het leven is vurrukkulluk	–	–	11	11
Zwartboek	3	5	13	21
Total	12	8	77	97

Table 4 – Distribution of fluency errors per error subtype

One of the main reasons to categorise a mistake in a sentence as a grammar error was incorrect word order. Not surprisingly, the incorrect word order in the Dutch TT was frequently similar to the word order in the English ST. Matamala and Ortiz-Boix (2016), who analysed the machine translation of AD in the Spanish-Catalan language pair, found that most annotated mistakes in their corpus (39.8%) belonged to this type as well.

11	Blind	2-44	EN (original)	At the same time his fingers try to pinpoint the branch wherever it lands.
			NL (MT)	Tegelijkertijd proberen zijn vingers te lokaliseren de tak, waar hij ook landt.
			NL (post-edited MT)	Tegelijkertijd proberen zijn vingers de tak te lokaliseren, waar hij ook landt

MT is prone to other grammatical errors as well. For instance, errors regarding the subject and verb agreement as in example 12, where the Dutch verb “open” (he opens) is in the second person singular where it should be plural.

12	Zwartboek	76	EN (original)	SS men on the boat open fire and Rob is hit in the chest.
			NL (MT)	SS-ers op de boot open vuur en Rob wordt in de borst geraakt.
			NL (post-edited MT)	SS’ers op de boot openen vuur en Rob wordt in de borst geraakt.

Another example are missing or incorrect prepositions, as in example 13, where the necessary Dutch preposition "van" (from) is missing or example 14, where the wrong Duch preposition "bij" was used instead of "langs".

13	Blind	1-27	EN (original)	Marie looks up from reading and sees Catherine staring at her as she sips a cup of tea, her little finger raised in a genteel fashion.
			NL (MT)	Marie kijkt op van het lezen en ziet Catherine naar haar staren terwijl ze nipt een kopje thee, haar vingertje op een vriendelijke manier opgetild.
			NL (post-edited MT)	Marie kijkt op van het lezen en ziet Catherine naar haar staren terwijl ze nipt van een kopje thee, haar vingertje op een deftige manier opgetild.
14	Zwartboek	79	EN (original)	Rachel swims through the darkness by the bank of the river.
			NL (MT)	Rachel zwemt door de duisternis bij de oever van de rivier.
			NL (post-edited MT)	Rachel zwemt door de duisternis langs de oever van de rivier.

Another example is wrong articles or anaphora. In Example 15, the Duch article "de" is used instead of "het", while in example 16 the article "het" is used while it should be "hem".

15	Blind	2-27	EN (original)	Marie suddenly closes the book, rises and leaves the room.
			NL (MT)	Marie sluit plotseling de boek, stijgt op en verlaat de kamer.
			NL (post-edited MT)	Marie sluit plotseling het boek, staat op en verlaat de kamer.
16	Blind	2-76	EN (original)	She offers him a towel. When he doesn't take it she puts it in his hand.
			NL (MT)	Zij biedt hem een handdoek aan. Als hij het niet aanneemt, legt ze het in zijn hand.

			NL (post-edited MT)	Zij biedt hem een handdoek aan. Wanneer hij hem niet aanneemt, legt ze hem in zijn hand.

A final example is the incorrect use of the possessive case (example 17).

17	Zwartboek	98	EN (original)	He gets up and shakes Rachel's hand.
			NL (MT)	Hij staat op en schudt Rachel's hand.
			NL (post-edited MT)	Hij staat op en schudt Rachels hand.

After wrong word order, wrong agreement was the dominant error category in the research of Matamala and Ortiz-Boix (2016). However, in this study, wrong articles and prepositions were categorized as wrong agreement as well. Moreover, Matamala and Ortiz-Boix' study (2016) focused on a language pair where grammatical gender is more pronounced than in the English-Dutch language pair.

4.3. Style and Other

Two smaller error types are Style and Other, with 44 and 24 errors respectively. The Style error category contains the *awkward* and *unidiomatic translation* error subtypes, which were reported by the evaluators to be hard to distinguish and often rather subjective. The concatenation of two identical possessive structures in example 18 is not necessarily incorrect, disregarding the spelling errors, but is unusual and awkward in Dutch. A combination of different possessive structures, e.g., a possessive 's combined with "van" (English: "of") + owner, would be more suitable here in Dutch. The unidiomatic translation in example 19 is a correct albeit literal translation of "judge his reaction". However, the Dutch language has an idiomatic verb plus noun combination to express this, which is why "beoordelen" has been replaced by "peilen" in the post-edited sentence.

18	Zwartboek	82	EN (original)	Rachel's father's wallet is taken and some gold ingots concealed in her mother's clothing are taken too.
			NL (MT)	Rachel's vader's portemonnee wordt meegenomen en er worden ook wat gouden staven in de kleding van haar moeder meegenomen.
			NL (post-edited MT)	De portemonnee van Rachels vader wordt meegenomen en er worden ook wat goudstaven uit de kleding van haar moeder meegenomen.

19	Blind	2-29	EN (original)	Marie looks at Ruben to judge his reaction.
			NL (MT)	Marie kijkt naar Ruben om zijn reactie te beoordelen.
			NL (post-edited MT)	Marie kijkt naar Ruben om zijn reactie te peilen.

We grouped the subtypes *mistake in source* and *unnecessary translation* in Other, the error category for miscellaneous errors. *Mistake in source* contains errors caused by the translation of an incorrect word in the source text. In example 20, "gets up" was misspelled in the original English text as "gest up", which was then translated to Dutch based on the English word "gestures", leading to a translation error as a result of a mistake in the source. *Unnecessary translations* were prevalent in the opening credits, where names, (broadcasting) companies and international terminology were part of the AD script. In example 21, this occurred with the erroneous translation of the name "Marina" as "Jachthaven", a Dutch word used to refer to a yacht wharf.

20	Zwartboek	56	EN (original)	He smiles and gest up from his desk.
			NL (MT)	Hij glimlacht en gebaart vanaf zijn bureau.
			EN (back translation)	He smiles and gestures from his desk.
21	Het leven is vurrukkulluk	20	EN (original)	Coproducent NTR Marina Blok
			NL (MT)	Coproducent NTR Jachthaven Blok
			EN (back translation)	Coproducer NTR Yacht-basin Blok

5. Conclusion

The present article contributes to research on the machine translation of AD. An overview of the existing literature clearly shows that limited research in this field has been undertaken to date, but that it may prove very useful in supporting the future development of this practice. AD is a text type with unique features, such as the multimodal nature of the source text and the intersemiotic dimension underlying the initial translation of that text. The impact of such characteristics on the machine translation output of AD is yet to be explored in detail. The present case study has highlighted several relevant issues that could form a basis to stimulate further research in this area.

A preliminary observation is that whereas the idiosyncratic language of AD with its relatively short and simple sentences was initially thought to be a good candidate for NMT, this case study has brought to light several challenges. First, the NMT output

demonstrates a significant error rate: 520 marked errors on a target text of 6,374 words, with 69.7% of all translated AD blocks containing one or more errors and an average of 2.3 errors per AD block. This points to a clear need for post-editing to bring the text up to the required quality standards. The extent to which post-editing is necessary and the effort required to correct the mistakes will have to be studied through experimental research, but the present case study already highlights that the multimodal context in which the translation is used might impact and potentially complicate the post-editing process. Certain mistakes are related to the multimodal context of the AD and can only be determined when also consulting the original images. The extent to which and how the original images are/can be consulted in the AD post-editing process, still needs to be studied.

A second observation is that the present study offers a first glimpse into the types of errors that occur most frequently. Particularly the categories of Accuracy/Mistranslation and Fluency/Grammar were most noticeable. As mentioned in the analysis, the main types of identified errors seem to suggest that in the course of the translation process, the NMT system often does not or cannot take into account the immediate context sufficiently enough and as a result it fails to arrive at adequate translations. Research into context-aware models for neural machine translation is progressing (see for example Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita, et al., 2018). The present case-study identified various errors that can be attributed to a lack of context for the translation, leading to mistranslations (by misinterpreting the multiple meanings of homonyms, for instance) and fluency errors (such as verb tense misinterpretations, incorrect anaphora, no subject-verb agreement and errors in word order). Further research will have to look whether such problems can be solved when more context-aware models are integrated in the NMT systems. Furthermore, in the case of AD the multimodal context in which the translation operates (sound effects and dialogues) and the original images of the film on which the AD is based need to be taken into account. As the case study indicates, not all mistranslations seem incorrect at first glance. A specific type of mistranslation is mistranslation due to the multimodal context, which is an extra dimension of context that the MT cannot (yet) consider in the translation process. The MT may output a correct sentence (cf. *Panda puts the photo on top of the cabinet - Panda legt de foto bovenop de kast)*, but the visual content that is part of the source text is necessary to distinguish the exact meaning, and consequently the correct translation of the sentence, in this case “zet” (used in Dutch when something is put somewhere vertically) rather than “legt” (used in Dutch when something is put somewhere horizontally). Previous studies also mentioned in this respect the translation of cultural references (see section 2). While this was not a frequent issue in the present case study (which might be due to the limited size of the sample), we did encounter some examples of translation errors of cultural references due to the lack of multimodal context. In this respect, researchers and developers are exploring the new domain of Multimodal Machine Translation (see Sulubacak et al. (2020) for a recent overview), which could potentially prove very useful to improve MT performance for multimodal text types such as AD since these systems would be able to take into account the multimodal context.

A third observation is that some errors might demonstrate that the system used has not been trained for AD specifically. Several types of errors result from the occurrence of AD-specific linguistic constructions (such as the frequent use of the present participle to express simultaneity, which research has shown is a typical feature of AD language; Reviers, 2018a; Salway, 2007). Particularly in the language pair English-Dutch, the use of the present participle and "as" to express simultaneity pose problems for MT. In Dutch speech and (non-AD) texts for general purposes, the use of the present participle is rather uncommon, which in this case study had a significant influence on the NMT engine and its performance when translating English AD into Dutch. An interesting area for further research, therefore, would be the development of MT systems trained for AD and an evaluation of the extent to which this improves the MT's performance.

Finally, previous studies also underlined the issues of timing and sentence length. Some languages might result in longer translations than the original. In AD, however, this can become problematic when the translated AD no longer fits into the pauses between dialogue and sound effects. This issue did not occur in this case study, which might be due to the limited sample size or indicate that this issue is less apparent in the English-Dutch language combination, compared to other languages, such as English and Spanish.

While the present article points to several interesting avenues for further research, it is only a preliminary case study with various limitations. First, the same error-type analyses should be conducted on larger samples and compared with other language pairs to corroborate the results. In addition, this case study included fiction drama films only, and the MT performance may differ for other types (documentary, corporate videos) and other genres (horror, comedy).

In addition, MT systems could be compared to human translations to identify not only differences in the types of errors but also in terms of style and norms. As scholars have mentioned (see section 2), AD norms and guidelines differ across countries and languages and as a result an MT translation might be correct, but not acceptable to the target audience. An analysis of the MT output's acceptability with target users is therefore a crucial consideration. Finally, more studies need to be conducted on measuring the actual post editing effort, in particular for the language pair English-Dutch, to complement existing findings and evaluate the actual effort and the impact of the multimodal context. The case study presented in this article was a pilot to a four-year PhD project in Translation Studies, funded by the University of Antwerp (2020-2024), in which these issues will be explored in more detail. In the first phase, the project will analyse the types of errors in a corpus of Dutch ADs, comparing the MT with the human translations of the same text. In the second phase, an experiment will be conducted to measure and compare the post-editing effort with translation of the original AD. This way we hope to shed more light on this new form of machine translation that will grow as more and more countries will be obliged to make more audio-visual content accessible to people with sight loss.

References

Álvarez Muniain, A.; Balenciaga, M.; Pozo Echezarreta, A. D.; Arzelus Irazusta, H.; Matamala, A.; Martínez Hinarejos, C. D. (2016). Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association (ELRA), pp. 3049-3053. <https://aclanthology.org/L16-1487/>. [Accessed: 20211116].

Arma, S. (2011). *The Language of filmic audio description*. PhD dissertation. Napoli: Universita Degli Studi di Napoli Federico II.

Bahdanau, D.; Cho, K. H.; Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Paper presented at 3rd International Conference on Learning Representations, ICLR 2015, San Diego.

Banerjee, S.; Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization.* Association for Computational Linguistics, pp. 65-72. <https://aclanthology.org/W05-0909/>. [Accessed: 20211116].

Bawden, R.; Sennrich, R.; Birch, A.; Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 1304-1313. <https://doi.org/10.18653/v1/N18-1118>. [Accessed: 20211116].

Bourne, J.; Hurtado, 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: Diaz-Cintas, J.; Orero, P.; Remael, A. (eds.). *Media for all: Subtitling for the Deaf, Audio Description, and Sign Language.* Amsterdam: Rodopi.

Bryssinck, J. (2019). *The (semi) automatic translation of audio descriptions: an exploratory study.* [MA thesis]. University of Antwerp.

Bywood, L.; Georgakopoulou, P.; Etchegoyhen, T. (2017). Embracing the threat: machine translation as a solution for subtitling. *Perspectives: Studies in Translation Theory and Practice,* v. 25, n. 3, pp. 492–508. <https://doi.org/10.1080/0907676X.2017.1291695>. [Accessed: 20211116].

Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds.). *Translation quality assessment: From principles to practice*. Cham: Springer, pp. 9-38. <https://doi.org/10.1007/978-3-319-91241-7_2>. [Accessed: 20211116].

Castilho, S.; Moorkens, J.; Gaspari, F.; Sennrich, R.; Way, A.; Georgakopoulou, P. (2018). Evaluating MT for massive open online courses. *Machine Translation,* v. 32, n. 3, pp. 255–278. <https://doi.org/10.1007/s10590-018-9221-y>. [Accessed: 20211116].

Cho, K.; Van Merrienboer, B.; Gulcehre, C.; Bougares, F.; Schwenk, H.; Bengio, Y. (2014a). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP). Association for Computational Linguistics, pp. 1724–1734. <https://doi.org/10.3115/v1/D14-1179>. [Accessed: 20211116].

Del Pozo, A. (2014). *Sumat: An Online Service for Subtitling by Machine Translation.* Sumat website. <http://www.fp7-sumat-project.eu/>. [Accessed: 20211116].

Etchegoyhen, T.; Bywood, L.; Fishel, M.; Georgakopoulou, P.; Jiang, J.; Van Loenhout, G.; Del Pozo, A.; Sepesy Maucec, M.; Turner, A.; Volk, M. (2014). Machine translation for subtitling: A large-scale evaluation. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC14).* European Language Resources Association, pp. 46–53. <https://aclanthology.org/L14-1392/>. [Accessed: 20211116].

Fernández-Torné, A. (2016). Machine translation evaluation through post-editing measures in audio description. *InTRAlinea: online translation journal* v. 18. <https://www.intralinea.org/index.php/archive/article/machine_translation_evaluation_through_post_editing_measures_in_audio_descr>. [Accessed: 20211116].

Fernández-Torné, A.; Matamala, A. (2016). Machine translation in audio description? Comparing creation, translation and post-editing efforts. *SKASE Journal of Translation and Interpretation,* v. 9, n. 1, pp. 64–87. <http://www.skase.sk/JTI10index.html>. [Accessed: 20211116].

Guerberof-Arenas, A.; Toral, A. (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces,* v. 9, n. 2, pp. 255–282. <https://doi.org/10.1075/ts.20035.gue>. [Accessed: 20211116].

Herrador Molina, D. (2006). *La Traducción de guiones de audiodescripción del inglés al español*. [Proyecto fin de carrera]. Universidad de Granada.

Jankowska, A. (2015). *Translating audio description scripts: translation as a new strategy of creating audio description.* London [etc.]: Peter Lang.

Jankowska, A.; Milc, M.; Fryer, L. (2017). Translating audio description scripts… into English. *SKASE Journal of Translation and Interpretation,* v. 10, n. 2, pp. 2–16. <http://www.skase.sk/JTI13index.html>. [Accessed: 20211116].

Liu, Y.; Tor-Carroggio, I.; Rovira-Esteva, S.; Casas-Tost, H. (2021). *Localisation guidelines for translating AD from Spanish into Chinese: A first proposal.* [Paper presentation]. Advanced research seminar on audio description, January 26-27.

Lommel, A. (2018). Metrics for translation quality assessment: A case for standardising error typologies. In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds.). *Translation quality assessment: From principles to practice.* Cham: Springer, pp. 109-127. <https://doi.org/10.1007/978-3-319-91241-7_6>. [Accessed: 20211116].

Lopez Vera, J. F. (2006). Translating Audio Description Scripts: The Way Forward?:Tentative First Stage Project Results. In: *MuTra 2006: Audiovisual*

Translation Scenarios Conference Proceedings: Audiovisual Translation Scenarios, pp. 1-10. <https://www.euroconferences.info/proceedings/2006_Proceedings/2006_proceedings.html>. [Accessed: 20211116].

Mälzer, N.; Schaeffer-Lacroix, E. (2021). The French-German TADS project: A multi method approach to the translation of AD scripts [Paper presentation]. *Advanced research seminar on audio description, online, January 26-27*. <https://grupsderecerca.uab.cat/arsad/node/149>. [Accessed: 20211116].

Macken, L.; Fonteyne, M.; Tezcan, A.; Daems, J. (2020). Assessing the comprehensibility of automatic translations (ArisToCAT). In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT2020).* European Association for Machine Translation, pp. 485-486. <https://aclanthology.org/2020.eamt-1.64/>. [Accessed: 20211116].

Matamala, A.; Ortiz- Boix, C. (2016). Accessibility and multilingualism: an exploratory study on the machine translation of audio descriptions. *TRANS,* n. 20, pp. 11–24. <https://doi.org/10.24310/TRANS.2016.v0i20.2059>. [Accessed: 20211116].

Matusov, E.; Wilken, P.; Georgakopoulou, P. (2019). Customizing neural machine translation for subtitling. In: *Proceedings of the Fourth Conference on Machine Translation: Vol* 1., Research Papers. Association for Computational Linguistics, pp. 82-93. <https://doi.org/10.18653/v1/W19-5209>. [Accessed: 20211116].

O'Hagan, M. (2019). *The Routledge Handbook of Translation and Technology*. London [etc.]: Routledge. <https://doi.org/10.4324/9781315311258>. [Accessed: 20211116].

O'Hagan, M. (2020). Introduction: Translation and technology: disruptive entanglement of human and machine. In: O'Hagan, M. (ed.). *The Routledge Handbook of Translation and Technology*. London [etc.]: Routledge, pp. 1-8. <https://doi.org/10.4324/9781315311258-1>. [Accessed: 20211116].

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (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. <https://doi.org/10.3115/1073083.1073135>. [Accessed: 20211116].

Pedersen, J. (2008). High felicity: A speech act approach to quality assessment in subtitling. In: Chiaro, D.; Heiss, C.; Bucaria C. *Between text and image: updating research in screen translation.* Amsterdam: John Benjamins, pp. 101-115. <https://doi.org/10.1075/btl.78.11ped>. [Accessed: 20211116].

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation.* Association for Computational Linguistics, pp. 392-395. <https://doi.org/10.18653/v1/W15-3049>. [Accessed: 20211116].

Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds.). *Translation*

quality assessment: From principles to practice. Cham: Springer, pp. 129-158. <https://doi.org/10.1007/978-3-319-91241-7_7>. [Accessed: 20211116].

Remael, A.; Vercauteren, G. (2010). The translation of recorded audio description from English into Dutch. *Perspectives: Studies in Translation Theory and Practice,* v. 18, n. 3, pp. 155–171. <https://doi.org/10.1080/0907676X.2010.485684>. [Accessed: 20211116].

Reviers, N. (2016). Audio description services in Europe: an update. *JoSTrans, The Journal of Specialised Translation,* n. 26, pp. 232-247. <https://www.jostrans.org/issue26/art_reviers.pdf>. [Accessed: 20211116].

Reviers, N. (2018a). *Audio description in Dutch: A corpus-based study into the linguistic features of a new, multimodal text type.* (PhD Book). University of Antwerp. <https://explore.lib.uliege.be/discovery/fulldisplay/alma9919943602602321/32ULG_INST:MOSA>. [Accessed: 20211116].

Reviers, N. (2018b). Tracking multimodal cohesion in Audio Description: Examples from a Dutch audio description corpus. *Linguistica Antverpiensia, New Series: Themes in Translation Studies,* v. 17, pp. 22-35. <https://doi.org/10.52034/lanstts.v17i0>. [Accessed: 20211116].

Salway, A. (2004). AuDesc System Specification and Prototypes. *TIWO: Television in Words.* University of Surrey. <https://andrewsalway.files.wordpress.com/2020/02/tiwo_television_in_words_deliverable_3-1.pdf>. [Accessed: 20211116].

Salway, A. (2007). A corpus based analysis of audio description. In: Diaz-Cintas, J.; Orero, P.; Remael, A. (eds.). *Media for all: Subtitling for the deaf, Audio Description and Sign Language.* Amsterdam [etc.]: Rodopi, pp. 151-174. <https://doi.org/10.1163/9789401209564_012>. [Accessed: 20211116].

Shterionov, D.; Superbo, R.; Nagle, P.; Casanellas, L.; O'Dowd, T.; Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation,* v. 32, n. 3, pp. 217–235. <https://doi.org/10.1007/s10590-018-9220-z>. [Accessed: 20211116].

Sulubacak, U.; Caglayan, O.; Grönroos, S.A.; Rouhe, A.; Elliot, D.; Specia, L.; Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation,* v. 34, pp. 97–147. <https://doi.org/10.1007/s10590-020-09250-0>. [Accessed: 20211116].

TAUS. (2021). *Harmonized DQF-MQM Error Typology.* <https://www.taus.net/qt21-project#harmonized-error-typology>. [Accessed: 20211116].

Tiedemann, J.; Scherrer, Y. (2017). Neural machine translation with extended context. In: *Proceedings of the Third Workshop on Discourse in Machine Translation (DISCOMT'17).* Association for Computational Linguistics, pp. 82-92. <https://doi.org/10.18653/v1/W17- 4811>. [Accessed: 20211116].

Toral, A.; Way, A. (2018). What level of quality can neural machine translation attain on literary text? In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds.). *Translation*

quality assessment: From principles to practice. Cham: Springer, pp. 263-287. <https://doi.org/10.1007/978-3-319-91241-7_12>. [Accessed: 20211116].

Uiterwijk, G. (2019). *The (semi) automatic translation of audio descriptions: an exploratory study.* [MA thesis] University of Antwerp.

Vercauteren, G. (2007). Towards a European guideline for audio description. In: Díaz-Cintas, J.; Orero, P.; Remael, A. (dds.). *Media for All: Subtitling for the Deaf, Audio Description, and Sign Language* (pp. 139–150). Amsterdam: Rodopi, pp. 139-150. <https://doi.org/10.1163/9789401209564_011>. [Accessed: 20211116].

Voita, E.; Serdyukov, P.; Sennrich, R.; Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers).* Association for Computational Linguistics, pp. 1264-1274. <https://doi.org/10.18653/v1/P18-1117>. [Accessed: 20211116].

Volk, M.; Sennrich, R.; Hardmeier, C.; Tidström, F. (2010). *Machine translation of TV subtitles for large scale production.* Paper presented at the 2nd joint EM+/CNGL Workshop Bringing MT to the user: research on integrating MT in the translation industry. Denver, Colorado.

Filmography

Van den Dop, T. (dir.) (2007). *Blind*. Netherlands, Belgium, Bulgaria.

Verhoeven, P. (dir.) (2006). *Zwartboek*. Netherlands, Germany, UK, Belgium.

Weisz, F. (dir.) (2018). *Het leven is vurrukkulluk*. Netherlands.