Journal of Audiovisual Translation

Insights From Mental Model Theory and Cognitive Narratology as a Tool for Content Selection in Audio Description

Gert Vercauteren[✉]
University of Antwerp

Citation: Vercauteren, G. (2021). Insights from Mental Model Theory and Cognitive Narratology as a Tool for Content Selection in Audio Description. *Journal of Audiovisual Translation, 4*(3), 6–24. https://doi.org/10.47476/jat.v4i3.2021.191

Editor(s): H. Silvester & T. Touminen

Received: May 25, 2021

Accepted: November 23, 2021

Published: December 28, 2021

Abstract

One of the main questions in audio description (AD) to which no systematic answers have been provided yet, is how to decide what information you include in your description and – if there is not enough time to describe everything – how you prioritize that information. In the present paper I want to propose an answer to this problem by asking the question: how do audiences process (filmic) stories and what information do they need to process them? The basic idea underlying this question is that people process and interpret stories by creating mental models (Johnson-Laird, 1983) of these stories. The paper explains how these models are created, what information is necessary to create them and what is optional, thus helping describers to decide what information in their description is "need-to-have" and what is "nice-to-have". The theoretical explanation will be applied to the opening of the film *Slumdog millionaire* (Boyle, 2008), to illustrate how the theory works and can be used in daily practice.

Key words: audio description, cognitive narratology, mental models, content selection.

[✉] gert.vercauteren@uantwerpen.be, https://orcid.org/0000-0001-6711-2005

1. Introduction

“A thousand words leave not the same deep impression as does a single deed.” This quote by Henrik Ibsen that was later paraphrased as “A picture is worth a thousand words,” must be the most used adage in audio description (AD) practice and research. Anyone who has ever described a film, TV series or any other audiovisual product knows how frustrating it can be to see all those marvellous, detailed images and not find the right words to adequately describe them or – probably even more frustrating – not have enough time to describe the visual wealth they are presented with. Therefore, it comes as no surprise that the question of content selection or what to describe is one of the topics that have received the most attention both in AD guidelines and AD research. In reply to the question of what must be described, the German guidelines for AD to film by Benecke & Dosch (2004) state that the answer is very easy, namely everything there is to see. A similar answer can be found in Neves (2011), who says that *in theory* everything that is presented on screen in visual terms should be included in the description. However, she goes on to mitigate that statement pointing out that in practice it is impossible and even undesirable to describe in words all the details that can be found in an image (p. 49). The question immediately arising from this observation is: what do you describe when you cannot include everything in your AD? It is precisely this question that will be the focus of the present paper. Based on insights from cognitive narratology, I will try to formulate an answer to that question that can help audio describers to select the most relevant content for their AD.

1.1. Content Selection in Existing AD Guidelines

In their comparison of AD guidelines, Greening et al. (2010) write that deciding “what to describe – what to include – what not to include – in what order – and how to prioritize the given information” (p. 4) is one of the most difficult problems that an audio describer is faced with. When it comes to the possible content of the description, all guidelines seem to be in agreement that this can be broken down into four main components: “when, where, who and what must be part of a description” (Greening et al., 2010, p. 5). The German guidelines created by Norddeutscher Rundfunk (2019) narrow down this general guidance by adding that these four main components should be clear to the target audience in any situation that is important for the course of action. In addition, sounds that cannot be identified without an additional explanation and any text on screen, including subtitles, signs and written messages, and credits, should be included as well (Greening et al., 2010, p. 5). On the one hand, this guideline is very general and vague, and in many instances it will be impossible to include all the elements mentioned. The question of how to prioritize the information to be included when time constraints force the audio describer to make a selection, is hardly ever treated at length in the guidelines. The most useful advice can be found in the German guidelines by Benecke and Dosch (2004), who state that the AD must include where the action is taking place and what characters are present, in addition to a description of the action itself. If after that there is still time, additional information can be given about clothing, furniture, colours and smaller, secondary actions (Benecke & Dosch, 2004, p. 21). Although this guideline does indeed allow describers to make a selection, it is clear that it is pretty anecdotic: what about emotions, scenery, etc.? Also, more concrete guidance could be given on how to distinguish between the main action and secondary actions. So far, the most comprehensive answer to what to describe and how to select the most relevant information for a description has been provided by the set of international guidelines (Remael et al., 2015) that resulted from the ADLAB PRO project[1]. The main difference between these guidelines and their predecessors, is that a) they are based on a theoretical foundation that allows for a systematic analysis and decision-making process and b) they advocate the use of strategies rather than hard-and-fast rules. They are also the only ones to not only look at what content is present in the original film or other audiovisual product, but also at what content the audience needs in order to understand the story that is being told. The idea that audience members are not just passive consumers but actively process and interpret the film or TV series they are watching and thus participate in the creation of its meaning, can offer a solid approach to decide which information to include and which to omit, as will be explained in section 2. It offers audio describers a systematic tool to prioritize information and allows them to explain why they made certain choices in their description, an important requirement in both instructional and professional contexts (Nord, 1991; Vercauteren, 2014).

1.2. Content Selection in Existing AD Research

While most of the AD guidelines discussed in the previous section are the work of practitioners, academics have looked at content selection in audio description too, and various approaches have been suggested. Di Giovanni (2014) builds on earlier work by Orero and Vilaró (2012) and Kruger (2012), and starts from the premise that viewers do not scan images randomly but use both automatic and controlled cognitive mechanisms to find the most relevant information in these images and to retrieve their meaning. She conducted an eye-tracking experiment to see what elements attract most attention of sighted viewers and used the results to create an audio description based on the findings of the experiment. Next, both the audio description created on the basis of eye-tracking and an earlier version were presented to a visually impaired audience, whose response to both versions was then compared in terms of comprehension and reception. It was found that the AD based on the eye-tracking data resulted in a better understanding than the earlier version, which points to the fact that eye-tracking could help with content selection in AD. While this approach clearly has the advantage of providing detailed and objective data with regard to what people find most relevant in a film, the main drawback seems to be its practical feasibility. Gathering eye-tracking data and processing them is a very time-consuming process that is not realistic in a professional context where descriptions have to be provided in ever shorter timespans.

Another approach adopted in studies on content selection in AD is corpus analysis. The basic principle underlying this line of research (e.g., Jiménez Hurtado, 2007, 2013; Salway, 2007; Reviers, 2015,

[1] EC Project Number: 2016-1-IT02-KA203-024311. Website: www.adlabproject.eu

2018) is that AD uses a specific language and that the study of that language can provide insights into various grammatical, stylistic and lexical features of that language, together with the kind of information that is included in audio descriptions. This line of research does provide valuable input as to what words and word categories are encountered most in AD and could therefore be considered as relevant; it may not, however, be the best candidate to give guidance on content selection for two reasons. First of all, by looking at specific words and word categories, corpus research focuses more on how ADs are formulated than on what specific content they contain. Second, they provide insight into how ADs have been written in the past without any indications as to their quality. This means that findings from corpus research in terms of described content will not automatically lead to reliable guidance for future descriptions.

A third line of research starts from the observation that many audiovisual products, including but not limited to films and TV series, tell a story, and that insights into how these stories are created (i.e., what content elements they contain) and into how they are processed and interpreted by the audience (i.e., what content elements the audience needs to understand the story that is being told), can help audio describers to select and prioritize the most relevant content. It is this narratological approach, explored amongst other academics by Pujol and Orero (2007), Kruger (2009, 2010); Kruger and Orero (2010), Vercauteren (2012, 2014); Vercauteren and Remael (2015) that will be adopted in the remainder of this paper.

2. Cognitive Narratology and Audio Description

As I already indicated in section 1.1., stories are not constructs that originate in the minds of their authors to be exclusively their work. For any story to work and truly come alive, active involvement on the part of the audience is required: people need to process and interpret the multiple cues that are presented to them to understand and enjoy the story (Herman, 2002, 2013). They do so by creating a mental model of that story, or a *storyworld*, to use Herman's (2002) term, who describes storyworlds as mental representations "of who did what to and with whom, where, why and in what fashion" (Herman, 2002, p. 5). This mental model creation is a very complex process that operates on various levels and involves different kinds of knowledge, but as will be explained in the following sections, even basic insights into how this process works can be beneficial for audio describers. If they know how mental model creation works and what information the audience really needs ("need to have") and what information is optional ("nice to have"), they can make informed decisions about how to prioritize and select the most relevant information for their description.

2.1 Mental Model Creation – a General Outline

Mental model theory is a theory developed by Johnson-Laird (1983). It posits that people make sense of the world around them by constructing mental representations of the specific situations in which they find themselves. Broadly speaking, these mental representations are created on the basis of two types of information: a) concrete information from the situation itself, used for the so-called *bottom-up processing* and b) personal, general knowledge, used for the so-called *top-down processing*. This mental modelling approach can also be applied to story processing. When the audience see Jamal Malik in the film *Slumdog millionaire* (Boyle, 2008) they will use the visual information on screen in a bottom-up process to create a mental representation of Jamal as a young man in his late teens or early twenties, with pitch black hair and brown eyes. In addition, aural information will be stored in the mental representation to later recognize his voice. As the story proceeds, this mental representation will continuously be updated and expanded: new information will be added, information that is no longer valid will be deleted and possibly replaced, and links will be created between this representation and others (e.g., between Jamal and the girl Latika, or between Jamal and the game show).

At the same time, story processing involves a *top-down* dimension in which the audience uses personal knowledge for at least three different reasons:

a) to add implicit and/or general information to the mental representations they create: in the opening of the film, we learn that Jamal grew up in a slum, so his mental representation can include the information that he was not very rich as a kid;

b) to make assumptions and create expectations about past and future events in the story. Since Jamal was not very rich, we may assume that he takes part in the game show *Who wants to be a millionaire* to win a lot of money and we may hope that he does win;

c) to make inferences about the events and entities that are referred to (both explicitly and implicitly) in the narrative. During the introduction of Jamal on the game show, there is a shot in which a young girl (Latika) is shown, who – as suggested by the camera work – appears to be looking up at him. Although this is not explicitly mentioned, we may infer that Jamal and Latika are related in some way.

How this mental modelling process works in the case of audio description has been explored by Braun (2007, 2016), Fresno (2016) and Fresno et al. (2016) among others. Braun (2007, 2016) takes a holistic approach and explores how mental modelling theory, originally developed for (verbal) discourse processing, can be applied to multimodal products such as films. She looks at the different semiotic channels that are used to create coherent mental models of films and shows where visually impaired viewers may have difficulties in maintaining that coherence. Based on these findings, she explains how audio describers can guarantee that this multimodal coherence is maintained in AD. Fresno (2016) focuses on the creation of mental models of narrative characters. She draws on insights from cognitive narratology, film studies and social psychology to theoretically explain how narrative film characters are (re)created by visually impaired audiences. Related research by Fresno et al. (2016) experimentally explores what character traits in audio description are best remembered by the audience, as a first indication of what information should get priority in the AD.

2.1.1. Types of Knowledge Used in Mental Model Creation

Throughout this continuous mental modelling process, various kinds of knowledge are used. In this section I will focus on two of these kinds of knowledge that provide most of the information for the mental model, namely *general knowledge* and *text-specific knowledge* (Emmott, 1997). The first type of knowledge, i.e., *general knowledge*, is knowledge that people acquire through their experiences and interaction with the real world (Emmott, 1997, p. 23). It is often referred to as *schemata,* defined by Emmott and Alexander (2014) as “cognitive structures representing generic knowledge, i.e. structures which do not contain information about particular entities, instances or events, but rather about their general form”. These schemata can be general representations of people or objects, such as a general mental representation of “a police officer” or “a dining room,” or general representations of events, such as a general mental representation of “a restaurant visit.” In story processing, it is this kind of general information that helps audiences make inferences about information that is not explicitly present in the text and helps them fill in gaps in the story that are so obvious that they do not have to be mentioned: when we see a man taking a shower in one scene and we see him having breakfast in the next, as in the opening scene from the film *Derailed* (Håfström, 2005), we know he must have washed, dried himself off, got dressed and walked to the kitchen. All these intermediate actions do not have to be shown explicitly. Similarly, this kind of general knowledge allows us to link objects together. When an audience sees a hand putting a kettle on a stove, based on the information stored in the schema of “a kitchen,” they will infer that this event is taking place in a kitchen.

The importance of the general information stored in schemata is obvious in the context of audio description, since it means that AD does not have to explicitly include a significant portion of information which can be derived from the schema it refers to. However, two observations have to be made with regard to the use of general knowledge and schemata in audio description.

First, it has to be clear that a schema is always a general, prototypical instance of the object, entity or situation it refers to, so describers have to gauge to what extent the specific instantiation of the schema shown on screen corresponds to the prototypical mental representation. For most of us the schema of “a hotel room” represents the idea of a relatively small room with a single or double bed, a closet, a television set and a separate room with a toilet and a shower. So, this is probably the image that will be conjured up when a description reads “In a hotel room”. But this image does not correspond at all to the hotel room shown in the opening scene of *The queen’s gambit* (Frank, 2020), namely a luxurious suite with a separate entry hall, high ceilings, vases with colourful flowers on marble columns and a spacious sitting area with baroque furniture. In other words, if the representation of a schema that is shown on screen resembles the prototype closely (in this case a “normal” hotel room), a basic description will be enough to create a mental model of it. If, on the other hand, the concrete representation does not resemble the prototype (as in the case of the suite), a much more elaborate description will be needed to create a suitable mental model.

Second, despite being called "general," general knowledge also comprises a more personal dimension. Take the description of the opening scene of the film *21* (Luketic, 2008): "Two narrow racing boats power along the river Charles". Most of us will be able to create a mental picture of this event, but only those who know where the river Charles runs, will be able to name the location where the event is set, namely Boston. This level of specificity in naming geographical, but also historical and other cultural references, has to be taken duly into account when creating a description. More general descriptions such as "From a meeting room window, a man looks at the office buildings in the distance" allow everybody to create a mental representation, albeit a very general one from which very little additional information can be inferred, e.g., in which country or city this action is set. A very specific description such as "From a meeting room window, a man looks at the Gherkin in the distance", will allow the audience to create a much more specific representation, setting the action in London's financial district, but there is a significant risk that some audience members will not be able to create a representation at all if they don't know what the Gherkin is. In this case, it may even have the undesired (comical?) effect of creating a wrong representation.

While general knowledge is essential for the processing and interpreting of stories, it is clear that in itself, it is not sufficient for a full understanding of any particular story. As its name already indicates, this type of knowledge only provides general and more or less prototypical information, but in no way does it provide any information that makes the story that is being told, unique. In other words, when we know that Jamal Malik is "a young man" we can derive certain general characteristics from that knowledge, but we still do not know anything about his physical appearance or about his personality traits. That is why *text-specific information* is equally important in the creation of a mental model of the storyworld: while the schemata deliver a blank blueprint, text-specific information is needed to fill these blueprints with concrete information. In fact, given their nature as mental containers of prototypical information, schemata cannot be used to store specific information about the story. Instead, audiences create separate containers for all characters and settings they encounter in the story, named *character representations* and *location representations*, respectively. These are pre-filled with general information that comes from the schema and is then supplemented with text-specific information in order to create a unique entity representation.

2.1.2. Processing Narrative Characters and Settings

From Herman's (2002) definition of a storyworld as a mental representation of "who did what to whom, where, why and in what fashion" (p. 5) it is clear that characters and the actions they perform and undergo, occupy a central position in the mental model creation process. Except from the "where" in this definition, all other questions relate to the characters, what they do, how they do and why they do it. In other words, character representations are fundamental in story processing, and as such also in AD. They basically consist of three types of information, namely information about the character's *physical properties,* about their *communicative and behavioural properties* and about their *mental properties* (Margolin, 2007; Vercauteren, 2012). The physical properties are the most stable ones: they refer to a character's external features such as age, gender, looks, etc. Since these features are usually relatively permanent, they constitute a relatively fixed dimension in the character representation and only need to be updated if they change. So, when Jamal's physical appearance is known it will remain in the character representation unaltered until, for example, he changes his clothing style. A character's *communicative and behavioural dimension*, as the name already suggests, gives information about the character's interactions – verbal and otherwise – with other characters in the story. These actions and reactions are closely linked to the third set of properties mentioned above, namely the character's *mental properties*. These refer to whatever a character sees, hears or feels, to their emotions and their wishes and desires. The audience uses information about this mental dimension of the characters to make sense of their interactions, i.e., to understand why they do something (e.g., a desire to reach a certain goal) or react in a certain way (e.g., because they feel sad or angry), to make assumptions about what has happened to the characters in the past and to form hypotheses about how they will behave in the future. From all this, it is clear that character representations are highly dynamic constructs. As explained by Schneider (2001), when a character is first presented in the story, the audience will create a new character representation using two initial strategies: on the one hand they will look for information that allows them to assign the new character to a certain category (cf. top-down processing using schemata), a strategy he calls *categorization*. Since audiences usually have limited time to get to know characters in films, films often rely on prototypical or even stereotypical characters, e.g., the good or the bad cop, so that a fair amount of general information about a character can be introduced through representation based on our general knowledge. At the same time audiences look for information that makes the character unique (cf. bottom-up processing using information from the text), such as the way they look, the clothes they wear, etc., a strategy Schneider (2001) calls *personalization*. When that character later reappears, the audience will look for information to further personalize it, and for information that validates the categorization that was made before. If this information is found, the categorization is confirmed. If not, there will be a temporary uncertainty, until information is presented that confirms the categorization, or signals that a character belongs to a different category. Most of the information that is used in this process comes from the character's communicative and behavioural dimension, complemented by its mental dimension.

For audio describers, this dynamic process means that they have to monitor characters, their (re)actions and their inner life throughout the entire film. When a character is first presented, they will have to include physical information in their AD to allow the audience to personalize the character. Since this information is stable (cf. before), once it has been mentioned in the AD, it does not need to be repeated until it changes. So, the describer can focus on the behavioural and mental information in the remainder of the AD, to allow the audience to assign the character to a particular category, to confirm this categorization or change it when contradictory information is presented, just like the sighted audience will.

The second type of representations created in the mental model are the so-called *location representations*, containing information about the different settings of the story. These settings comprise a temporal dimension indicating when the action is taking place and a spatial dimension indicating where it is taking place (Vercauteren & Remael, 2015), and they are an essential component of any story. As Herman and Vervaeck (2005) state: “story development is inconceivable without the setting, which makes it possible for actions to take place and actants to become involved in them” (p. 56). For a long time, however, the study of settings received very little attention in narratology. Dennerlein (2009) attributes this to the fact that the study of narrative was initially focused on the construction of the story, on the representation and the causal sequence of events (p. 4), and Herman (2002) writes that “if space was discussed at all it was used negatively to mark off setting from story…” (p. 265). Later, it was shown that settings are much more than merely the background for the events they were first believed to be (see for example Dennerlein, 2009; Pitkänen, 2003 or Vercauteren & Remael, 2015 for an account specifically focused on setting and AD). Two elements are of particular relevance for content selection in AD. First of all, settings can and often do have a symbolic function that has to be taken duly into account when describing them and that will often require more description than when they just provide background. For example, the cubicles where Jamal and his colleagues work are the same colour as the workers’ shirts, meaning that the workers are not more important than the environment in which they work. Second, settings are often linked to specific characters in the story. In *The Hours* (Daldry, 2002), Virginia Woolf is linked to Richmond in 1923, Laura Brown is linked to Los Angeles in 1951 and Clarissa Vaughan is linked to New York in 2001. In other words, as soon as this link is created in the description, and hence in the audience’s mental model, the describer can just refer to one of the elements, e.g., Clarissa, New York or 2001, to conjure up the entire setting, which means there will be more time to describe other elements. This being said, research in cognitive narratology has shown that the creation of location representations in mental models is problematic in various respects. In an experiment Ryan (2003) asked her students to draw the setting of a short story she read out loud to them, and found out that the mental reconstruction of a setting is a very personal process that can differ significantly from one person to the next. This seems to confirm what Emmott (1997) wrote earlier, namely that a lot of detailed information about settings in stories tends to be forgotten (p. 38). Moreover Dannenberg (2008) found that detailed location representations are not needed for the audience to become immersed in the story (p. 85). Indeed, Dannenberg (2008) and Herman (2002, 2009) found that, for processing narrative settings, audiences predominantly use basic schemata that deal with their static dimension on the one hand (i.e., the contents of the setting or “what it looks like”) and with their dynamic dimension (i.e., possibilities to move within and beyond a setting and links or “paths” between settings). Although more research is needed here, this seems to suggest that detailed spatial information is not a prerequisite in AD and describers can give priority to other types of information if they are pressed for time.

2.3. Putting It All Together

As became apparent in the previous sections, mental model creation is a continuous process. Moreover, all the elements described above, i.e., the use of general world knowledge and text-specific knowledge, the creation of character representations and of location representations, the tracking of the development of events, combine to create one overall mental representation of the story. Building on the work of Emmott (1997), who used mental model theory to explain how audiences interpret personal pronouns, the following paragraphs explain how this process works for narratives as a whole and how it can be applied to audio description to determine and decide what to include in the AD and what to leave out. The opening scene of the film *Slumdog millionaire* (Boyle, 2008) will be used as an example to show how this approach works in practice.

The basic premise underlying the entire process is that when audiences are presented with an event in a story, they make a mental representation of it and, in order to do so, they need to create a general context for that event, called the *fictional context* (Emmott, 1997, p. 103). This is stored in a *contextual frame* that is created on the basis of information from the text and inferences made from that information, and at a minimum it consists of a) the characters that are present at this event, b) the location where the event takes place and c) the global timeframe within which it takes place. It is important to point out that contextual frames only consist of general information and the so-called *episodic information*, i.e., information that is linked to that particular event or situation in the story but not necessarily to other events or situations. For example, the film *Slumdog millionaire* (Boyle, 2008) opens with a heavy-handed interrogation. The contextual frame for that event consists of the characters that are present there, namely Jamal Malik and a police officer, the location of the interrogation, namely a police cell somewhere in Mumbai and a global time indication of when the interrogation takes place, namely 2006. This very general contextual information is episodic in that it only refers to this event and not to the next one. Indeed, in the next scene Jamal is introduced to the gameshow *Who Wants to Be a Millionaire*. So, although Jamal is still present as a character in the contextual frame of the second event, the police officer is not, and another character, i.e. the show's host, is present instead. The event is set in a TV studio and based on the fact that Jamal looks exactly like his representation in the first event, we can infer that this event also takes place in 2006; this piece of information, however, is not explicitly presented and we have to infer it. So, since the event is different, episodic information that relates to the gameshow will be stored in a different contextual frame.

In addition to general and episodic information, the audience are also presented with a lot of data that are true beyond the immediate context of these events: the audience learn what Jamal, the police officer and the show host look like, they get detailed information about the interrogation room and the studio where the game show is set, etc. The details that are independent of any specific event, are what Emmott (1997) calls *non-episodic information* (p. 122) and they are stored in the so-called *entity representations* that can take the form of character representations storing detailed information about characters and location representations storing detailed information about settings (Emmott, 1997, p. 104). Based on the example presented above, we can conclude that there is no general way to distinguish between what is episodic and non-episodic information, since every event is unique and thus presents highly specific details. Information that is non-episodic in one film (e.g., a character's hair colour) may be episodic in another and only a careful analysis of the film will allow the describer to decide what (episodic) information is part of the contextual frame and what (non-episodic) information is part of the entity representations.

As the story progresses, the audience will continuously monitor new developments and evolutions to keep track of the (chrono)logical chain of events, try and motivate characters' actions and reactions, hypothesize about future events, etc., basically by paying close attention to what stays the same and what changes. This process is what Emmott (1997) calls *contextual monitoring* (p. 112). It starts with the creation of links between the different entities that are present in a certain contextual frame: in the example of *Slumdog millionaire* (Boyle, 2008) above, Jamal and the police officer are linked to each other, to the interrogation room and to Mumbai, 2006. Likewise, Jamal and the host are linked to each other, to the TV studio and (presumably) to Mumbai, 2006. This process is called *binding* (Emmott, 1997, p. 122) and is based on the essential principle that entities remain bound to a certain contextual frame until the story tells otherwise. In other words, if the story told in *Slumdog millionaire* (Boyle, 2008) returns to the interrogation room after the scene in the TV studio, the audience will assume Jamal and the police officer are still there. Likewise, if the audience again hear the sound of the show host, they will assume the story returned to the game show and Jamal will also be there. This principle is crucial in mental model creation and contextual monitoring because it significantly reduces the processing load imposed on the audience, who do not have to check continuously whether all the entities of the frame are still there and can assume that they are unless signalled otherwise. It is clear that this principle is equally crucial for AD: by mentioning one of the entities that are bound to a certain frame, the description will conjure up the entire context, which does not have to be described in all detail every time. So, an audio description reading “Back in the TV studio” not only tells us where we are, but also tells us who is present and when the events take place; a very valuable advantage for audio describers who are often pressed for time.

In addition to this basic principle of binding, Emmott (1997) distinguishes between three different scenarios that can occur when moving from one scene in the story to the next. First of all, characters can enter or leave a frame that is being monitored, a scenario that is called *frame modification* (Emmott, 1997, p. 142). So, while the audience continues to monitor the same contextual frame, its composition changes. For AD this means that the describer will have to signal who has left the frame and who has entered it, in other words, what characters are unbound from the frame and what new characters are bound to it. For these new characters, new general “labels” will be created in the contextual frame and individual entity representations will be created to store non-episodic, specific details that make those characters unique. A second possible scenario occurs when the audience stop monitoring a certain frame and start monitoring a different one; this is called *frame switch* (Emmott, 1997, p. 147). This scenario clearly requires more processing effort by the audience, since it involves a change of location, possibly combined with a change in time. In addition, it often also comprises a change of configuration of the characters that are present in the frame. It is evident that this scenario also requires more audio description. Describers have to mention the new location and time, and indicate what characters are present, so that their target audience can create the new general contextual frame. In addition, if time allows, more detailed information can be given to fill the new, lower-level location and character representations. The third scenario explained by Emmott, is the *frame recall* (Emmott, 1997, p. 150). As the name already indicates, this scenario happens when the story returns to a frame that was presented earlier. It is in this case that the binding principle explained above is particularly useful. As Emmott (1997) points out: “when a frame is re-primed, it is not necessary to mention each element of the frame again. A mention of one element means that the other, being bound to it, can be re-primed automatically.” (p. 152). In other words, when *Slumdog millionaire* (Boyle, 2008) returns to the interrogation after the first scene in the TV studio, an image of either the interrogation room or the police officer is enough to recall the entire contextual frame that is linked to the interrogation. This is interesting for AD as well, as was already explained above. If after the scene in the TV studio the AD for example mentions “The police officer” to indicate who is speaking, the audience knows that we are back in the interrogation room in Mumbai, 2006 and Jamal is also there. This possibility to recall a contextual frame by mentioning just one element in it, is very convenient when there is little time for description, but it also gives the audio describer more time to add information to the entity representations if there is more time.

The diagram below shows a schematic representation of how this content selection process can take place in practice.

Figure 1

Analysis Form for Content Selection in AD

	SCENE 1	
Action(s) / Event(s):		
	CONTEXTUAL MONITORING / FRAME	
CHARACTERS ‘Label’ / Name	**TIME** Global (and/or Local) Time	**SPACE** Frame
	ENTITY REPRESENTATIONS	
CHARACTERS Physical Psychological Communicative/Behavioural	**TIME** Local Time Weather Conditions	**SPACE** Contents Relations

SEQUENTIAL ↓ **MONITORING**

	SCENE 2	
Action(s) / Event(s):		
	CONTEXTUAL MONITORING • Modification • Switch • Recall	
CHARACTERS ‘Label’ / Name	**TIME** Global (and/or Local) Time	**SPACE** Frame
	ENTITY REPRESENTATIONS	
CHARACTERS Physical Psychological Communicative/Behavioural	**TIME** Local Time Weather Conditions	**SPACE** Contents Relations

Source: author’s own work

When an audio describer is presented with a film to describe, starting with the first scene (s)he will first look at the events/actions taking place. As explained above, these form the backbone of the mental model of the narrative and as such they will have to be included in the AD; otherwise, the audience will not be able to start the creation of their mental model. Next, based on the theory put forward by Emmott (1997), the describer will look for the contextual frame of these events/actions, i.e., for general, non-episodic information on the characters that are present in the scene and on the time and space where the events/actions are set. For all these general contextual elements, specific character and location representations will be created in which detailed, non-episodic information is stored. After this first analytical step, the describer can then decide what information to include in the AD, depending on the available time and after a careful analysis of what information can be derived from the other semiotic channels such as dialogues or sound effects, starting from the general (necessary) information and moving on to the more detailed, non-episodic information if time allows.

When the scene changes, this process will be repeated with a few specific additions and adaptations. First, the describer will check what happened to the contextual frame with the scene change. There can be a frame modification in which the setting remains the same, but characters leave the frame and new characters enter, for which new contextual "labels" and character representations will have to be created. There can also be a frame switch in which the action moves to a different setting (with the same or different characters). In this case a new contextual frame will have to be created, together with new character and location representations. As from the third scene onwards, the third scenario, i.e., a frame recall, also becomes possible. In that case, existing contextual information is re-primed and the describer can add new information to character and location representations if this is presented in the frame that is being recalled. I am very well aware that this is a fairly theoretical description. To make it more concrete, a practical illustration, using the opening scene from *Slumdog millionaire* (Boyle, 2008) discussed above, can be found in Appendix 1. In addition, it is clear that it may be impractical to use this scheme while working since it may be very time consuming to analyse every scene in this way. However, it may serve a didactic function as it can be used by audio description trainees who are still learning to select relevant content for their descriptions and have more time at hand to practice this process.

3. Conclusion

While there is already some research on "what to describe," the question of how to prioritize information that is eligible for description has received far less attention. It is, however, equally important since very often describers will not have time to include everything they want in their description and they will have to make a selection. In the present paper I explored the possibility of using insights from mental model theory and cognitive narratology for content selection in AD. I explained how audiences who are presented with a narrative (audiovisual) text process and interpret it by creating a mental model in which they use both general knowledge (particularly in the form of schemata) and knowledge they gain from the text. While these schemata can be used to provide general containers in the mental model, information from the text is necessary to fill these containers (or representations) with specific narrative information to make the mental model concrete. Based on the model developed by Emmott (1997), I explained that actions or events form the backbone of the mental model and that audiences process and interpret them by adding two levels of information to them: episodic, contextual information in the form of general "labels" for the characters who perform and undergo the actions and for the settings in which they take place, and non-episodic character and setting information that is stored in dedicated character and location representations. Based on the example from *Slumdog millionaire* (Boyle, 2008) that was used to illustrate this mental modelling approach, it has become clear that this approach can also be used to determine which information to include in a specific AD and what to omit, based on what should get priority in the mental model at that time. However, this was only the first exploration of this approach. Various questions remain, pointing at different possible avenues for further research. First of all, the approach presented here has so far only been described in theory and its usefulness in practice will have to be tested. As already indicated in the last section, it may serve first and foremost a didactic purpose for training the beginning audio describers and further research may test whether or not it helps them a) choose relevant content and b) motivate their choices. A second, more fundamental limitation of the model presented here is that – so far – it only looks at the different visual channels in the audiovisual source text and does not explain how the aural channels (music, sound effects and dialogues) contribute to the creation of the mental model and hence to the prioritization of information for AD. It has been pointed out before that sound has been under researched, both in narratology (Mildorf & Kinzel, 2016) and in AD (Remael, 2012; Szarkowska & Orero, 2014), and better understanding of how audiences (both sighted and blind) use sound to process and interpret stories, of how audio describers use sound to create their audio descriptions and of how sounds and audio descriptions interact to create meaning and re-create th0e story that is told in the original, is urgently called for and would form a valuable and, in fact, an indispensable addition to the model presented in this paper. Finally, the difference between need-to-have contextual information and nice-to-have entity information can offer a valuable starting point to prioritize content in AD, but more research is needed to determine further prioritization criteria for the information contained in the entity representations. Both the images shown on screen and other information such as dialogues and sound effects will co-determine what to choose, but as Ryan (2003) has shown, the processing of narrative space is a very individual endeavour and a lot of information that is presented in the narrative gets lost in its reconstruction by the audience. And as Schneider (2001) illustrates, what character information is necessary and optional strongly depends on the specific position of a character at any given point in the narrative. The consequences for these findings for content selection and prioritization in AD have yet to be explored.

References

Benecke, B., & Dosch, E. (2004). Wenn aus Bildern Worte werden. [When pictures become words]. Bayerischer Rundfunk.

Boyle, D. (Director) (2008). Slumdog millionaire [Film]. Celador Films; Film4; Fox Searchlight Pictures; Warner Bros.; Pathé.

Braun, S. (2007). Audio description from a discourse perspective. A socially relevant framework for research and training. Linguistica Antverpiensia New Series, 6, 357–369.

Braun, S. (2016). The importance of being relevant? A cognitive-pragmatic framework for conceptualising audiovisual translation. Target, 28(2), 302–313. https://doi.org/10.1075/target.28.2.10bra

Daldry, S. (2002). The hours [Film]. Paramount Pictures; Miramax; Scott Rudin Productions.

Dannenberg, H. (2008). Coincidence and counterfactuality. Plotting time and space in narrative fiction. Lincoln.

Dennerlein, K. (2009). Narratologie des Raumes [Narratology of Space]. De Gruyter.

Di Giovanni, E. (2014). Visual and narrative priorities of the blind and non-blind: Eye tracking and audio description. Perspectives. Studies in Translation Theory and Practice, 22, 136–153. https://doi.org/10.1080/0907676X.2013.769610

Emmott, C. (1997). Narrative comprehension. A discourse perspective. Oxford University Press.

Emmott, C., & Alexander, M. (2014). Schemata. In P. E. A. Hühn (Ed.), The living handbook of narratology. Retrieved December 14, 2021 from https://www.lhn.uni-hamburg.de/node/33.html

Frank, S. (Executive Director) (2020). The queen's gambit [TV series]. Flitcraft; Wonderful Films; Netflix.

Fresno, N. (2016). Carving charachters in the mind. A theoretical approach to the reception of characters in audio described films. Hermeneutics, 18, 59–92.

Fresno, N., Castellà, J., & Soler-Vilageliu, O. (2016). 'What Should I Say?' Tentative criteria to prioritize information in the audio description of film characters. In A. Matamala & P. Orero (Eds.), Researching audio description (pp. 143–167). John Benjamins Publishing Company .

Greening, J., Petré, L., & Rai, S. (2010). A comparative study of audio description guidelines prevalent in different countries. RNIB.

Håfström, M. (Director) (2005). Derailed [Film]. Di Bonaventura Films; Miramax; Patalex V Productions Limited.

Herman, D. (2002). Story logic: Problems and possibilities of narrative. University of Nebraska Press.

Herman, D. (2009). Basic elements of narrative. Wiley-Blackwell.

Herman, D. (2013). Cognitive narratology. In P. E. A. Hühn (Ed.), The living handbook of narratology. Retrieved December 14, 2021 from https://www.lhn.uni-hamburg.de/node/33.html

Herman, L., & Vervaeck, B. (2005). Handbook of narrative analysis. University of Nebraska Press.

Jiménez Hurtado, C. (2007). La Audiodescripción desde la representación del conocimiento general. Configuración semántica de una gramática local del texto audiodescrito [Audio description as a representation of general knowledge. Semantic configuration of a local grammar of the audio described text]. Linguistica Antverpiensa, New Series: Themes in Translation Studies, 6, 345–356.

Jiménez Hurtado, C., & Soler Gallego, S. (2013). Multimodality, translation and accessibility: A corpus-based study of audio description. Perspectives. Studies in Translation Theory and Practice, 21(4), 577–594. https://doi.org/10.1080/0907676X.2013.831921

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference and consciousness. Cambridge University Press.

Kruger, J.-L. (2009). The translation of narrative fiction: Impostulating the narrative origo. Perspectives: Studies in Translatology, 17(1), 15–32.

Kruger, J.-L. (2010). Audio narration: Re-narrativising film. Perspectives: Studies in Translatology, 18(3), 231–249.

Kruger, J.-L. (2012). Making meaning in AVT: Eye tracking and viewer construction of narrative. Perspectives. Studies in Translation Theory and Practice, 20(1), 67–86. https://doi.org/10.1080/0907676X.2011.632688

Kruger, J.-L., & Orero, P. (2010). Introduction: Audio description, audio narration: A new era in AVT. Perspectives. Studies in Translation Theory and Practice, 18(3), 141–142. https://doi.org/10.1080/0907676X.2010.487664

Luketic, R. (Director) (2008). 21 [Film]. Columbia Pictures; Relativity Media; Trigger Street Productions; Michael De Luca Productions; GH Three.

Margolin, U. (2007). Character. In D. Herman (Ed.), The Cambridge companion to narrative (pp. 66–79). Cambridge University Press.

Mildorf, J., & Kinzel, T. (2016). Audionarratology. Prolegomena to a research paradigm exploring sound and narrative. In J. Mildorf & T. Kinzel (Eds.), Audionarratology. Interfaces of sound and narrative (Vol. 52, pp. 1–26). De Gruyter.

Neves, J. (2011). Guia de audiodescrição. Imagens que se ouvem [Audio description guide. Images that can be heard]. Leiria.

Nord, C. (1991). Text analysis in translation: Theory, methodology and didactic application of a model for translation-oriented text analysis. Rodopi.

Norddeutscher Rundfunk (2019). Forgaben für Audiodeskriptionen. https://www.ndr.de/fernsehen/barrierefreie_angebote/audiodeskription/Vorgaben-fuer-Audiodeskriptionen,audiodeskription140.html

Orero, P., & Vilaró, A. (2012). Eye tracking analysis of minor details in films for audio description. MonTI. Monografías de Traducción e Interpretación(4), 295–319. https://doi.org/10.6035/MonTI.2012.4.13

Pitkänen, K. (2003). The spatio-temporal setting in written narrative fiction [Unpublished doctoral dissertation]. University of Helsinki..

Pujol, J., & Orero, P. (2007). Audio description precursors: Ekphrasis, film narrators and radio journalists. Translation Watch Quarterly, 3(2), 49–60.

Remael, A. (2012). For the use of sound. Film sound analysis for audio description. Some key issues. Monti, 4, 255–276.

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

Reviers, N. (2015). The language of audio description in Dutch: Results of a corpus study. In A. Jankowska & A. Szarkowska (Eds.), New Points of View on Audiovisual Translation and Accessibility (pp. 167-189). Peter Lang.

Reviers, N. (2018). Tracking multimodal cohesion in audio description: Examples from a Dutch audio description corpus. Linguistica Antverpiensa, New Series: Themes in Translation Studies, 17, 22–35.

Ryan, M. L. (2003). Cognitive maps and the construction of narrative space. In D. Herman (Ed.), Narrative theory and the cognitive sciences (pp. 214–242). CSLI Publication.

Salway, A. (2007). A corpus based analysis of audio description. In J. Diaz-Cintas, P. Orero, & A. Remael (Eds.), Media for all: Subtitling for the deaf, audio description and sign language (pp. 151–174). Rodopi.

Schneider, R. (2001). Toward a cognitive theory of literary character. The dynamics of mental-model construction. Style, 35(4), 607–640.

Szarkowska, A., & Orero, P. (2014). The importance of sound for audio description. In A. Maszerowska, P. Orero, & A. Matamala (Eds.), Audio description: New perspectives illustrated (pp. 121–139). John Benjamins Publishing Company.

Vercauteren, G. (2012). A narratological approach to content selection in audio description. Towards a strategy for the description of narratological time. Monti, 4, 207–231.

Vercauteren, G. (2014). A translational and narratological approach to audio describing narrative characters. TTR: Traduction, Terminologie, Rédaction, 27(2), 71. https://doi.org/10.7202/1037746ar

Vercauteren, G. (2016). A translational and narratological approach to audio describing narrative characters. TTR, XXVII (2), 71–90.

Vercauteren, G., & Remael, A. (2015). Spatio-temporal settings. In A. Maszerowska, A. Matamala, & P. Orero (Eds.), Audio description. New perspectives illustrated (Vol. 112, pp. 61–80). John Benjamins Publishing Company.

Appendix 1.

Content selection form opening scene *Slumdog Millionaire* (Boyle, 2008)

SCENE 1		
Action(s) / Event(s): (Heavy-handed) Interrogation • Questions (verbal) • Blowing smoke in Jamal's eyes • Slapping Jamal in the face		
CONTEXTUAL MONITORING / FRAME		
CHARACTERS	**TIME**	**SPACE**
• Jamal • Police Officer	• 2006	• Mumbai • Interrogation room
ENTITY REPRESENTATIONS		
CHARACTERS	**TIME**	**SPACE**
Jamal • Young, black hair, stained white shirt • Scared, sweating, crying **Police officer** • Middle-Aged, Corpulent, bald, puffy eyes, khaki uniform • Angry, Unfriendly	Inside location with no access to the outside. No specific temporal details about this episode are known.	• Harsh yellow light • Cold, plastered walls • Wooden table, two chairs • No windows

CONTEXTUAL MONITORING

- Modification
- **Switch** ↓
- Recall

<table>
<tr><th colspan="3">SCENE 2</th></tr>
<tr><td colspan="3"><b>Action(s) / Event(s):</b> Start of the game show
• Pep talk by the host behind the scenes
• Appearance of Jamal and the host on stage
• Taking their seats</td></tr>
<tr><th>CHARACTERS</th><th>TIME</th><th>SPACE</th></tr>
<tr><td>• Jamal
• TV show’s host</td><td>• Not shown/mentioned</td><td>• Global space not shown/mentioned
• TV Studio</td></tr>
<tr><th colspan="3">ENTITY REPRESENTATIONS</th></tr>
<tr><th>CHARACTERS</th><th>TIME</th><th>SPACE</th></tr>
<tr><td><b><u>Jamal</u></b>
• Same entity as in previous contextual frame
• Stained shirt is still clean
<b><u>Host</u></b>
• Middle-aged, dark-haired, trimmed beard, dressed in a suit
• Extraverted, condescending</td><td>• Not mentioned/shown</td><td>• Dark behind the scenes
• Circular stage, metal frame, glass floor, metal stand, two metal stools in the centre
• Audience seated around the stage</td></tr>
</table>