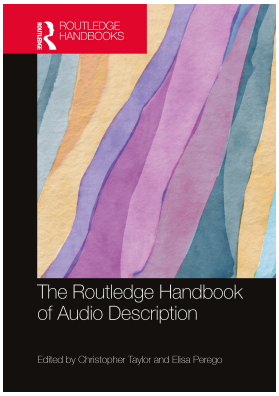

The Routledge Handbook of Audio Description

Christopher Taylor, Elisa Perego

A cognitive approach to audio description

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781003003052-7
Jana Holsanova
Published online on: 07 Apr 2022

4

A cognitive approach to audio description

Production and reception processes

Jana Holsanova

1. Introduction

Although audio description (AD) is a very young field of research, various aspects of AD have been studied within a number of disciplines, such as translation studies, discourse studies, film and media studies, accessibility and audiovisual translation studies (AVT), linguistics, psychology and narratology. Researchers have used various theories and methods and defined audio description from various perspectives. However, there is currently a great need for cognitively-oriented research on AD.

In accessibility studies, AD has been defined as a tool that provides access to audiovisual materials for people with visual impairments and blindness (BVI), offering a richer and more detailed understanding and experience of film, theatre or other events. In the field of translation studies, AD has been characterised as intermodal audiovisual translation, since the content from images is transferred into words (Jakobson, 1959; Reviers, 2017; Starr, 2017), and as partial translation, since it only translates parts of the film, while its users receive and process the other components, such as sounds, music and dialogues from the original production (Reviers, 2017). The research focus has mainly been on the mediating role of the audio describers: "audio describers must respect the communicative purpose of the film intended by the producer and adequately translate the message. At the same time, they must meet the communicative needs of the BVI recipients and adapt AD to it" (Reviers, 2017: 34).

Cognitive approaches to human communication focus on how we process and integrate information from different senses, how we perceive, attend to, understand and experience complex messages that others convey and how we remember these messages. In sum, the connection between language, thinking and meaning-making is central in this type of research. The cognitive approach to AD in particular concerns meaning-making processes during reception of the original material, selection and decision-making processes during production of AD and meaning-making processes during reception of AD by BVI audiences. Closest to the cognitive perspective on AD are therefore approaches that highlight it as a cognitive-linguistic meaning-making activity (Braun, 2007) and as a multimodal activity (Reviers, 2017).

In what follows, we will look more closely at perceptual and cognitive processes during the production and reception of audio described films. Film is conceived of as a *complex*

DOI: 10.4324/9781003003052-7

multimodal text, an organised arrangement of various semiotic resources that has been created in order to tell a story and achieve a certain communicative effect. Each of the resources (also called modes of expression) has a certain meaning and importance and contributes to the overall meaning of the text. According to Zabalbeascoa (2008), a filmic text consists of various configurations of four basic semiotic modes: (1) visual – non-verbal (e.g., images of objects, environments, characters, their gazes, gestures, mimics and body movements), (2) visual – verbal (e.g., text on the screen), (3) aural – verbal (e.g., dialogue) and (4) aural – non-verbal (e.g., music and sound effects). The film is both a result of the film production process and a starting point for reception processes since it offers a potential for meaning-making[1] (Kress and van Leeuwen, 1996).

Audio description can be studied both from a product perspective and from a process perspective. When regarding AD as a *product*, researchers study the results of the AD process *offline* (e.g., by analysing manuscripts containing verbal descriptions). When regarding AD as a *process*, researchers focus on audio describers' thought and meaning-making processes by tracing these activities *online*, during AD production. Even reception of AD by BVI audiences can be viewed from these two perspectives – either *offline*, as a reflection of the end-users' comprehension, experiences and memory of the film, or *online*, as a dynamic process of recipients' meaning-making, emotions and involvement during the performance (cf. section 3 Research Methods).

This chapter gives an overview of research, theories and methods used when investigating perceptual and cognitive processes underlying the production and reception of AD. In section 2, we highlight a number of research questions and issues that are important to investigate from a cognitive perspective. In section 3, we map out a variety of suitable methods that can be used when conducting empirical research and answering these research questions. In section 4, we review relevant theories that can be employed when studying production and reception of AD. Finally, in section 5, we formulate suggestions for future research.

2. Critical issues

In this section, we will look more closely at the various phases of audio description and highlight important research issues regarding cognitive processes, including how audio describers create meaning on the basis of an original film as a complex multimodal text, how audio describers decide what is relevant to describe at a particular moment in the film and how they verbalise it when producing AD. We will also look at how BVI audiences understand AD, how they mentally imagine described things and events and how they create meaning with the help of AD, in conjunction with the dialogues and sounds from the original film. Figure 4.1 illustrates these various phases in audio description.

The starting point is the original film as a complex multimodal text (*Production 1*, *P1*). The original film that has been created for sighted audiences has to be made accessible for BVI audiences by means of AD. Filmic text is a form of narration that makes use of visual scenes with objects, environments, characters, their gazes, gestures, mimics and body movements, as well as dialogue, text on the screen and music and sound effects. The audio describers are both recipients and producers. As sighted recipients of the multimodal text, they create meaning on the basis of the various expression modalities they perceive and interpret the main message of the film (*Reception 1*, *R1*). In the role of producers, the audio describers create an accessible version of a film for the BVI audiences by producing AD (*Production 2*, *P2*). These two activities are intertwined.

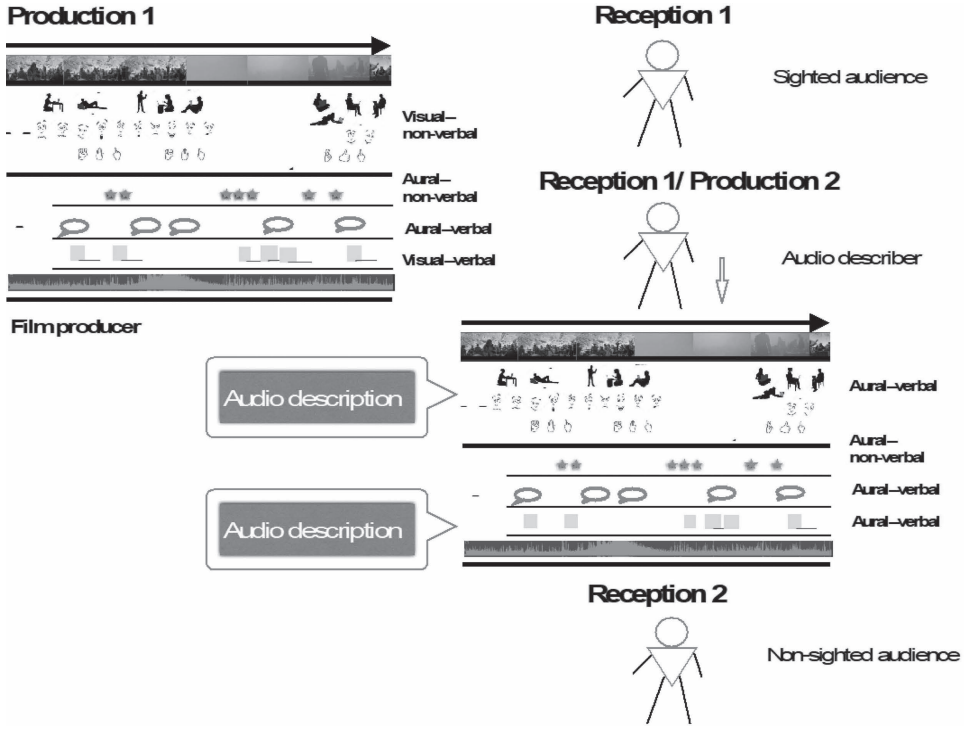

Figure 4.1 Overview of phases in AD production and reception (P1 – R1 – P2 – R2)

The AD is tailored and adapted to the needs and preferences of the target group. This means that the audio describers must consider the interplay of the modalities and examine which information is conveyed by the sounds, music and dialogues (and can be perceived by the BVI audiences) and which information is expressed only visually (and cannot be perceived by the BVI audiences). Based on this assessment, the audio describers then select relevant visual information and verbalise it. Since the interaction between the different modalities ceases to work when there is no access to the visuals, the audio describers attempt to restore this interaction through the verbalisation. In other words, the audio describers supplement what is missing from the multimodal interplay in order to achieve a comparable understanding and experience of the non-sighted audience (Reviers, 2017).

The AD is either produced live by an audio describer at a given venue or is pre-recorded. The audio described content is received and interpreted by the BVI audiences, along with the music, sounds and dialogues from the original film, either live, at the cinema, or via a DVD track or an application (*Reception 2*, *R2*).

In the following, we will look closer at the phases R1, P2 and R2, focus on cognitive and perceptual processes involved in these three phases and formulate critical issues and research questions that can be explored from a cognitive perspective. These three phases will also serve as a framework for the review of theoretical and methodological approaches to AD (sections 3 and 4).

Reception, *R1*. This phase concerns the process of multimodal text comprehension by the sighted audience in general and by the audio describer in particular. Reception is closely connected to the recipients' ability to select, attend to and process information, as well as to their ability to

integrate information from various resources and to fill in the "gaps" (Holsanova, 2014b). It is a rather complex process of deriving the content from the various modalities with the help of prior knowledge and making inferences on the basis of the context (Bucher, 2017; Wildfeuer, 2012). When watching film, audiences process both the visual, verbal and acoustic input and they integrate it into their previous world knowledge, filmic knowledge and mental representations, to reconstruct the story in their minds (Matamala & Remael, 2015). Reception of film stories is a very complex and multimodal process where the recipient must keep track of who did what to whom, where, when and why for each event in the narrative and constantly update this information according to how the story develops (Vercauteren & Remael, 2014; Zwaan & Radvansky, 1998).

The following questions must be explored: which activities are the audio describers involved in during the reception of the original multimodal text? How do they understand verbal, visual and auditory modalities and their interplay? What catches their attention and what does not? How do they find relevant information? How do they identify and interpret the main message? How do they – step by step – extract and integrate information from various resources in order to create meaning?

We still know very little about how people make sense of signs in a specific context, and the dynamic, processing perspective is still missing (Braun, 2007). Therefore, the focus of future research should be on empirical research on reception of multimodal messages and on the dynamics of multimodal meaning-making (Holsanova, 2014b). With respect to AD, the goal is to track these processes in audio describers in order to provide them with effective comprehension and interpretation strategies.

Production, *P2*. This phase concerns assessment, selection and decision-making processes during the production of AD (cf also Braun, 2007; Vercauteren, 2014). The audio describer selects relevant information from the visual scene (environment, events, people, their appearance, age, facial expressions, gestures and body movements), verbalises it by using vivid descriptions in a way that meets the needs of the blind and visually impaired audience (Holsanova, 2016a, 2016b). The audio descriptions activate inner images and perceptions in the end-users and make it easier for them to understand and experience what is happening (Johansson, 2016).

There are several challenges that audio describers struggle with. One of them is related to time constraints. The ideal response is to insert the description into the natural pauses and gaps in the dialogue so that it does not interfere with the dialogue, sounds and music (Benecke, 2004; Remael et al., 2014). This limitation influences the audio describers' decisions about what, when and how to describe. Another challenge is that language and images are two systems that work differently. Language is linear and the content is formulated sequentially, whereas images are analogue and function simultaneously. The richness of the visual mode is another challenge connected to this. Visual scenes and images contain a lot of information, but the AD cannot cover everything that can be seen. The question is which visual information should be left out? Furthermore, images enable interpretation at different levels of abstraction (Holsanova, 2011; Remael et al., 2016).

This brings us to the issue of relevance. Audio describers must assess which visual information is contextually essential for an understanding of events. As Braun puts it:

> In the overall flow of visual, verbal and auditory input, the audio describer needs to identify relevant cues and assess their contributions, i.e., evaluate them in terms of their relative weight (decisive, reinforcing, confirming, complementary, contradictory, etc.) and reliability at the point of occurrence.
>
> *(Braun, 2007: 5)*

Further challenges are connected to the level of detail in the description and to the choice of linguistic solutions to recreate visual meaning. Audio describers must prioritise information and carefully consider what they should include to achieve an effective AD that does not affect BVI audiences cognitively (Vercauteren, 2016).

More concretely, the following questions must be explored: how do the audio describers decompose the multimodal text and evaluate the semantic value of the visual, verbal and auditory input in order to single out the contribution of the visual cues and decide what to describe? How do they determine whether visual information is expressed by other modes of communication? How do they decide which part of the visual information is relevant and necessary for understanding the narrative? How do audio describers prioritise information that has to be mediated? How do they select appropriate verbal means for describing the visual content? How much information is needed? Which information must be mentioned explicitly and which can be inferred by the audiences?

It is important to systematically investigate the dynamics of meaning-making, selection and decision-making processes during the production of AD, in order to find effective strategies for what to describe, how to describe it, when to describe it and how to prioritise information. Since all the aforementioned processes in P2 imply a heavy burden on audio describers' mental resources, future research should focus on empirical studies testing various solutions in order to find optimal strategies for AD production.

Reception, *R2*. Reception of audio described films concerns the way BVI audiences process and understand AD, create meaning and enjoy and remember the story. Audio description is both about conveying the content so that target audiences do not miss important information that is only visible, and about mediating experience in order to elicit mental images and stimulate emotional responses in the target group. Thus, even the vocal delivery is important since both voice quality and prosody (rhythm, pace, intonation, emphasis, pausing) can carry meaning, highlight certain information and affect comprehension and enjoyment during reception (R2). In sum, the critical issues involved in the reception of AD concern information processing, comprehension, emotional response, engagement, immersion, mental imagery, cognitive load and memory.

In other words, the following questions must be explored: how do BVI audiences perceive, understand and experience linguistic descriptions of visual events? Are they able to follow and enjoy the story? Do they feel involved? Do the intended emotions come across? How do they imagine environments, characters and events? What kinds of descriptions do they prefer? What kinds of descriptions are most relevant to them? Do extensive descriptions lead to cognitive processing overload?

Cognitive reception studies are still very rare. More empirical research must be conducted to understand how the end-users process audio described film (Holsanova et al., 2020). Researchers can either investigate reception during an ongoing film performance – focusing on the dynamic process of the actual interaction with the film, or after a film performance – focusing on how the story was understood and remembered. Reception studies can also be conducted independently of a concrete performance with a focus on attitudes, viewing habits, general problems and preferences of the BVI audiences.

3. Research methods

In this section, we will map out a variety of suitable methods that can be used when conducting empirical research and answering questions and issues highlighted in the previous section. Since AD is a very complex phenomenon to study, it needs to be investigated from different

perspectives, using different methods and theoretical frameworks (Holsanova, 2016a). The majority of AD studies have been descriptive and *product-oriented*, using text linguistic or corpus linguistic methods to analyse AD manuscripts, namely, the outcomes of audio describers' meaning-making and decision-making processes. However, *process-oriented* studies that focus on the ongoing cognitive processes leading to these products are still rare. Also, much less attention has been devoted to study *reception* of AD, during or in connection to a performance.

In the following, we will review online and offline methods that can be used to uncover perceptual and cognitive processes during audio describers' reception of the original film (*R1*), during audio describers' production of AD (*P2*), as well as methods used to track reception of AD by the BVI audiences (*R2*).

Methods for tracing cognitive processes underlying audio describers' reception of film and production of AD

How can we trace the ongoing meaning-making activities that the audio describers activate during film reception? Partial knowledge about the meaning-making processes during film reception, as well as selection and decision-making processes during AD production, can be acquired indirectly, *offline*, by using interviews, for example. Vandaele (2012) mentions *introspection* as a method to come closer to one's own cognitive processes and mental states while engaging with film narrative.

However, in order to access the dynamics of such reception and production processes in detail, methods for *online* monitoring are more advantageous (e.g., think-aloud protocols, eye tracking, keystroke logging etc.). In addition, a combination of methods is sometimes required to get deeper insights into cognitive and interpretative processes.

Think-aloud protocol is a suitable method that enables researchers to come closer to thought processes. It is a qualitative research method where participants speak aloud when performing a task (Ericsson & Simon, 1993; van Someren et al., 1994). Think-aloud protocol is recorded online and then transcribed and analysed by the researcher. In general, this method is used to provide insights about participants' thinking and decision-making processes, especially regarding language-based activities. Applied to AD, the use of concurrent think-aloud protocols during the audio description task enables us to uncover steps in the dynamic meaning-making process regarding interaction with the multimodal text and the decision-making processes of AD production.

Jankowska (2019), for instance, adopts think-aloud protocols together with other online methods (eye tracking, keyboard logging and screen recording) to study the decision-making process of describers watching clips from Polish and Spanish films. Holsanova (2020b) uses data from audio description and think-aloud protocols during the AD task to uncover the dynamic interpretative processes of meaning-making in the interaction with a complex multimodal text.

Keystroke logging is a non-intrusive research method developed in order to collect online process data of writing activities used to study cognitive writing processes. Another online method suitable for the study of film reception and AD production is *eye tracking*. Eye tracking methodology allows for both explorative and experimental studies and can be used as an online process measure presenting a detailed account of the underlying perceptual and cognitive processes (van Gogh & Scheiter, 2009). The eye movements recorded by eye tracking software "provide an unobtrusive, sensitive, real-time behavioural index of ongoing visual and cognitive processing" (Henderson & Ferreira, 2004: 18). We can infer what attracts users'

attention and what does not, which elements have been attended to, in what order, for how long, how often and how carefully. In previous research, eye tracking methodology has been used to study the process of image viewing and image description, to reveal the underlying attentional processes (Holsanova, 2001, 2008, 2011) and to study reception of multimodality (for an overview cf. Holsanova, 2014a). When applied to AD, eye tracking of sighted viewers has been used to inform production of AD for the blind (Kruger, 2012; Di Giovanni, 2014). When keystroke logging is combined with eye tracking measurements, it gives the analyst an enhanced picture of the writer's attention processes: which objects or areas were scanned visually and which objects or areas were described in writing at a certain point of time (Andersson et al., 2006). Applied to AD, this combination of methods can capture thought processes while the audio describer is drafting an AD manuscript.

Since one of the main challenges of AD is to select relevant aspects of the visual scene from a large amount of visual detail, eye tracking studies have mainly focused on tracing information selection and focalisation via areas of interest that have been frequently fixated by sighted viewers. The problem is, however, that eye movement data alone do not tell us about recipients' understanding of the film narrative. We cannot conclude from the eye movement data alone what aspects and properties of the image have been focused on, why and at what level of abstraction. Visual fixation on an object does not reveal which concept was associated with it, or what the viewer had in mind. If we want to gain more insight into the audio describers' meaning-making processes, eye tracking must be combined with other methods (Holsanova, 2012, cf. also section 4). A possible solution is to use a dynamic, multimodal scoring method, by combining eye movement data and verbal protocol data (Holsanova, 2001, 2008), where the verbal protocol functions as a referential framework that enables researchers to infer the ideas and thoughts to which the visual fixations correspond. Another possibility is to combine eye movement data and narrative descriptions (Kruger, 2012).

All the aforementioned online methods are particularly useful for collecting process data in order to reveal the complex and creative process of meaning-making. They are important for a detailed study of perceptual and cognitive processes associated with the reception of film, the production of AD and for the identification of successful AD strategies.

Methods for tracing cognitive processes underlying reception of AD by BVI audiences

The traditional methods used in reception research are surveys, questionnaires, interviews and recall tests. These methods are used to reconstruct reception indirectly, offline, based on quantitative data. Chmiel and Mazur (2012) report a number of small and large-scale studies based on *questionnaires*. They present their own studies on users' comprehension, preferences, problems and suggestions and discuss a number of methodological issues. Among other useful suggestions, they recommend using questionnaires with concrete AD examples, instead of general questions asking about users' experiences only.

Another offline method is the holding of a *focus group* discussion. The forming of a focus group is a qualitative method used to collect the perceptions, opinions, beliefs and attitudes of a small group of carefully selected participants. Participants express their views about a specific product or discuss a specified issue under the guidance of a moderator. Applied to AD, a focus group can be used to obtain insight in various aspects of AD on the basis of a previously watched film with AD. It can either address the BVI target group only or a mixed group of sighted and BVI audiences, discussing extracts from authentic audio described films together (Holsanova, 2016b).

Holsanova et al. (2015) used a combination of methods to investigate BVI users' preferences in different genres. First, *interviews* with BVI users were conducted based on concrete AD examples. Second, alternative versions of audio described programmes were created, based on the variables extracted from previous interviews. Third, these alternative AD versions were evaluated by the end-users offline, (a) via quantitative *ratings* of informativeness, vividness, satisfaction and immersion and (b) via end-users' qualitative explanations during *focus group discussions*.

Based on results from previous studies and on assumptions from various theories, researchers can set up *experiments*. Experiments are conducted in order to support, refute, or validate an assumption and can either be hypotheses-driven (based on theories, e.g., cognitive load, memory, mental imagery, relevance), or data-driven (based on results from previous descriptive studies and corpus studies of AD). In the context of AD, Fresno et al. (2014) conducted an experimental study of users' recall of film character descriptions. This reception study was designed to test how the amount of information included in the audio description of characters and its presentation affect users' recall. The study was guided by two hypotheses. H1: due to memory limitations, the more information included in the AD, the more difficult its recall. H2: some strategies might help to reduce the extraneous cognitive load in the audio description of characters. In this study, the authors used memory span test and recall questionnaires as offline measures.

Yet another possibility is to conduct an *exploratory* reception study, which is carried out to test new ideas or applications. In the area of AD, for target groups beyond BVI audiences, Starr (2017) developed a tailor-made version of AD to test it on audiences with cognitive accessibility needs to investigate descriptions of affect, emotions and states of mind. In her study, Starr also included intervention and semi-structured interviews.

Studies using *online* methods to measure end-users' response during film perception are still extremely rare. Calderazzo (2010) suggests *skin conductance response* (SCR) to measure positive and negative emotions during reception of AD. Various stimuli (visual, auditive, olfactory) evoke time-related changes in skin conductance. SCR is based on recording variances in skin electrical conductivity as emotions fluctuate. When applied to audience response, this methodology allows the researchers to monitor audience's emotional reactions to various versions and styles of AD directly during film perception.

Another method to track direct response is (to use) *EEG methodology*. EEG, which stands for electroencephalography, is based on a non-invasive procedure of recording electrical activity from the scalp surface. It enables researchers to study and understand processes that underlie behaviour. In a pilot study, Kruger et al. (2017) used EEG methodology as an objective online measure, along with post-hoc questionnaires as a subjective offline tool, to measure the audiences' fluctuating degree of immersion in the story world.

The aforementioned online and offline methods for tracking perceptual and cognitive processes during reception of film and during production and reception of AD (R1, P2 and R2) are summarised in Table 4.1.

4. Current research: theories and concepts

Insights into perceptual and cognitive processes underlying production and reception of AD come from various research areas and theoretical perspectives. In the following, we will present relevant theoretical accounts and current research findings that can be employed when studying the three phases of AD (R1, P2 and R2) and the underlying cognitive processes.

Table 4.1 Summary of online and offline methods that can be used to track perceptual and cognitive processes during reception of film and production and reception of AD (R1, P2, R2)

Production/ Reception/ Methods	*Production 1 (P1) by the film-maker*	*Reception 1 (R1) by the sighted audio describer*	*Production 2 (P2) by the audio describer*	*Reception 2 (R2) by the BVI audience*
Process	Film production processes; selection and decision-making, orchestration of semiotic resources to achieve communicative effect	Meaning-making processes during perception and comprehension of the film	Meaning-making processes during production of the AD narrative; selection and decision-making processes	Meaning-making processes during the comprehension of the AD narrative, in conjunction with the original film sound and dialogues
Analysis and Evaluation Methods	N/A	*online methods:* eye tracking, think-aloud protocol, keystroke logging to measure perception, attention and thought processes	*online methods*: eye tracking, think-aloud protocol, keystroke logging to track visual attention, thought processes, selection and decision	*online methods*: skin conductance response to track cognitive load, emotions and involvement
Product	Film as a multimodal text	Interpreted version of the original film in the mind of the audio-describer and in notes	AD script/ recorded track with AD	Interpreted version of the audio described film = the AD
Analysis and Evaluation Methods	N/A	*offline evaluation methods*: studying notes, conducting interviews	*offline evaluation methods*: interviews with audio describers analysis: corpus studies, descriptive studies (manuscript as a product)	*offline evaluation methods*: ratings, interviews, focus groups or combination of methods to evaluate comprehension, memory and impressions after a concrete performance. Experiments: • hypotheses-driven (based on previous studies and theories); • data-driven (based on results from previous descriptive studies and corpus studies of AD). Surveys and interviews with BVI (independent of a concrete performance) to evaluate general attitudes, viewing habits, problems, preferences

Cognitive processes underlying audio describers' film reception and production of AD

An important role for meaning-making during film reception and for selection and decision-making processes during production of AD (R1 and P2), is played by theories on scene perception, visual attention, event segmentation, multimodality, relevance and information structure.

The consensus among researchers in *scene perception* (Henderson, 2007) is that visual perception and attention is guided by both low-level (bottom-up) processes and by high-level (top-down) processes. This means that where we look in a scene is partly determined by the scene constraints and driven by low-level image features, such as luminance, contrast, edge density, colour and motion (Itti & Koch, 2000) and partly by high-level factors, such as our expectations, interests, intentions, task, goal and previous knowledge. Visual perception is active, exploratory, creative and highly selective. Viewers filter what they perceive, select certain aspects of it and fill in other aspects on the basis of their expectations, knowledge and expertise (Holsanova, 2014b). Thus, one and the same scene can be viewed and conceptualised differently by different viewers. This has consequences for AD. It means that a scene description will include both features that several sighted viewers would agree on and features that will be coloured by high-level factors guiding the perception of the individual audio describer (Holsanova, 2016a).

So to what extent do sighted viewers agree on what is important in a visual scene? Do they always focus on the same objects at the same time? How similar are their scanpaths: can the viewing patterns of the sighted viewers be used as a guidance for audio describers indicating what to include in the scene description? Holsanova (2001, 2008, 2011) combined eye tracking and verbal description data and developed a dynamic, sequential method to study the process of image viewing and image description by sighted viewers. The eye movement data showed what image elements were attended to, when and for how long and the verbal description data showed how the image was actually perceived and experienced by the viewers. These two types of data were synchronised in the form of multimodal score sheets and their temporal and semantic relations were analysed. Although viewers agreed on a number of image elements in the first part of their visual discovery, their overall scanpaths differed due to the impact of high-level processes. Also, their image descriptions differed in style. While the more static, *descriptive* style focused on spatial arrangement and visual details, the *narrative* style focused on the temporal and dynamic aspects and on creating a coherent story.[2] In contrast to the consumption of static images, viewing dynamic images is more guided. Researchers in film psychology introduced the concept of *attentional synchrony* to name the act of directing the visual attention of the viewers to a specific element in a scene at a specific moment in time (Smith & Henderson, 2008). Among film techniques that can effectively guide visual attention are, for instance, sudden onset of motion, cuts to close-up, sharper focus on an object, higher contrast between light and dark areas of scenes and using off-screen sounds to direct audio attention to objects out of shot (Smith, 2013).[3] It has been shown that these techniques provide low-level, stimulus-driven control of attention and eliminate some of the individual differences based on high-level, motivation-driven control of attention. However, it still remains to be established to what degree film viewers' eye movements are affected by their higher-level narrative comprehension (Loschky et al., 2015).

Kruger (2012) takes a step in this direction. He makes a link between how sighted, hearing viewers receive and understand film (based on written narrative accounts) and how they look at film (based on eye tracking data). The findings suggest that "visually peripheral elements

that play a covert, top-down role in the narrative . . . gain particular narrative importance when competing with the more overt, bottom-up aspects of the narrative . . .” (Kruger, 2012: 67). Since audio describers often struggle with the choice of essential visual elements in the scene, this method could support them in their selection and decision-making processes. The ultimate goal is to create an effective AD that would “allow the blind viewer to construct a mental or imaginative interpretation of the story world” (ibid, p. 70). It is, however, still unclear whether the viewing priorities of sighted audiences correspond to those of blind and partially sighted individuals. Nevertheless, the combination of eye tracking methodology with spoken description or written narration is highly relevant for the research on AD since they provide windows on the dynamic processes of scene viewing and uncover the creativity and complexity of meaning-making.

Reception of film stories is a very complex cognitive multimodal process where the recipient must minimally keep track of who did what to whom, where, when and why for each event in the narrative and constantly update this information according to how the story develops (Vercauteren & Remael, 2014; Zwaan & Radvansky, 1998). According to Zwaan et al. (1995), sighted recipients construct coherent mental representations of the filmic situations and events in order to understand the film narrative. According to this theoretical account, people use models or *schemata* to this end, based on their personal life experiences, perceptions and understanding of the world, stored in long-term memory. Such schemata also provide a framework for acquiring and interpreting new information. During the reception of audio described films, such schematic knowledge is triggered and activated in the receivers’ minds by the keywords used in AD.

An important contribution to this line of research is a study by Fresno (2014) and Fresno et al. (2014) shedding light on the role of memory for reception, processing and comprehension of audio described film. Long term memory is a permanent store containing knowledge and experience acquired throughout life, whereas working memory has very limited capacity and enables the recipient to only briefly store and manipulate information. This has consequences for the production and reception of AD. Fresno et al. (2014) show that the amount of information included in the AD has an effect on reception and that segmented, stepwise descriptions are less cognitively demanding for the BVI audiences and lead to a better recall. The study was inspired by the principles of reducing *cognitive load*, formulated and tested in the area of multimedia learning (Chandler & Sweller, 1991; Mayer, 2005; see also Holsanova et al., 2009).

Event segmentation is another factor that plays an important role in the management of cognitive load and comprehension. Event segmentation concerns the human ability to conceive the boundaries of when an event starts and ends and is an important aspect of mental model construction (Radvansky & Zacks, 2014). All stories have temporal and spatial dimensions connected to their characters and the actions they are involved in. When watching movies, viewers segment information into meaningful events according to changes in the spatiotemporal reference frame of the film (Zwaan & Radvansky, 1998; Zwaan et al., 1995). Research shows that there is large overlap in what people perceive as distinct events, suggesting that there are cognitive principles for segmenting observed situations (Zacks et al., 2009). Event segmentation abilities are linked to the updating of working memory and to the integration of novel information with associated long-term memories (Radvansky & Zacks, 2014; Zwaan & Radvansky, 1998). If one segments well, this improves comprehension and saves valuable cognitive resources (Boltz, 1992). Event segmentation is therefore fundamental for comprehending and remembering narratives for both the sighted and the blind. To date, virtually nothing is known about how, or if at all, the sighted and the blind differ in how they perceive

and remember events. The question arises whether AD can contribute to film comprehension by explicitly marking event borders (Holsanova et al., 2020).

The area of *multimodality* research is relevant for the study of reception and production of AD since it uncovers how language, images and sounds jointly create meaning (Kress & van Leeuwen, 2001; Wildfeuer, 2012). Film has been characterised as multimodal text since it often uses multiple, overlapping modes of expression. This helps the audience to understand the meaning of a scene more easily, as several resources provide clues to its interpretation. Braun (2008: 5) argues that "the comprehension of multimodal texts is 'holistic', drawing on input from all modes involved". Audio describers are involved in a complex meaning-making activity and supplement what is lacking in the multimodal interplay and create content links between the AD, the sound and the dialogue so that the audience can reconstruct the story correctly and effectively (Reviers & Remael, 2015).

However, very few empirical studies have been conducted on processual aspects of multimodal meaning-making in general and AD in particular. Jankowska (2019) studies thought processes during production of AD by using eye tracking and other online methods. Cámara and Espasa (2011) focus on audio description of scientific multimedia presentations. The authors focus on scientific translation and accessibility, analyse existing audio described documentaries and propose alternatives that can improve visual accessibility of multimedia scientific texts. Holsanova (2020a) traces the interpretative process of meaning-making in users' interaction with complex popular scientific journal by combining data from audio description and concurrent think-aloud protocols. The study reveals how audio describers combine the contents of the available resources, make judgements about relevant information, determine ways of verbalising visual information, use conceptual knowledge, fill in the gaps missing in the interplay of the resources and reorder information for optimal flow and understanding.

The issue of deciding which aspects in the visual scene are important to include in AD is closely coupled to inferential models of communication and relevance theory (Grice, 1975; Wilson & Sperber, 2004). The central idea with the *Relevance Principle* is that communicators try to be as relevant as possible and the audience knows it and searches in any given communicative situation for the intended meaning. According to Wilson and Sperber (2004: 610), "human cognition tends to be geared to the maximisation of relevance." Communicators use various *ostensive stimuli* in order to attract an audience's attention to relevant information. Also, communicators try to convey useful information that would not cause mental effort for the audiences. What is expressed (verbally or non-verbally) creates expectations which guide the audience towards the intended meaning (Forceville, 2014). The intended meaning is then deduced from the input, from the context and with the help of background knowledge.

Applied to AD, there is the issue of what needs to be spelt out and what the audience is able to infer (see discussion in Braun, 2007 and Vercauteren, 2007). Palmer and Salway (2015) emphasise that communication and comprehension is inferential and goes beyond what can be seen. Kruger recommends to give priority to the "narrative implication or effect of what can be seen" (Kruger, 2010: 234). Some researchers suggest that AD should include descriptions of (implied) thoughts and emotions. Palmer and Salway (2015) propose in this vein that AD should not only include identification of characters and their actions but also knowledge about the mental states of the characters (e.g., their beliefs, desires and goals that motivate and explain the characters' actions). For instance, one and the same film scene can be described as either "They are *standing* behind the curtain", focusing on the action, or as "They are *hiding* behind the curtain", taking into account the context of the action and implying the mental states of the characters. This has consequences for both production and reception of AD and should therefore be evaluated by BVI audiences.

Concerning AD production and reception, *information structure* plays an important role. Information packaging is related to how elements are represented as either "known" or "new" information and how they are anchored in recipients' knowledge. Kluckhohn (2005) shows that linguistic information can be structured in a way that is optimally processed by the listener and that a particular word order can help to clarify scene changes. Vandaele (2012) suggests using concepts from cognitive linguistics and cognitive semantics in the production of AD, in order to trigger mental imagery in the audience. The suggested concepts include the activation of textual features such as landmark, trajectory and container, as well as the relations between figure-ground, source-goal, center-periphery, viewpoint, distance and direction of viewing (cf. Hirvonen, 2014 for analysis of figure and ground in filmic AD).

Cognitive processes underlying reception of AD by the BVI audiences (R2)

For the study of the reception of AD, focusing on how the BVI audiences perceive, imagine, understand, remember and enjoy the audio described film, the following areas of research are becoming increasingly relevant: theories on mental imagery, information processing, cognitive load, engagement, immersion and memory. Even the study of vocal delivery is important for the study of comprehension and enjoyment of AD.

According to Finke (1989: 2), *mental imagery* is "the mental invention or recreation of an experience that in at least some respects resembles the experience of actually perceiving an object or an event". Current research in cognitive science shows that mental imagery relies to a large degree on the same processes as those that are active during actual perception when we act upon the external world (Johansson et al., 2006, 2013; Laeng et al., 2014; Richardson et al., 2009). For instance, the brain is activated in the same way when we see a tree or imagine a tree, when we pick an apple or imagine picking an apple. In other words, the same activation occurs when we take in information through direct sensory input and when we create internal mental images in our minds. In connection to AD, the following questions arise: do both sighted and blind persons create mental images? What consequences does this have for production and reception of AD?

It has been shown that mental imagery in sighted persons is accompanied by spontaneous eye movements that closely reflect the content and spatial layout of the imagined scene (Johansson et al., 2006, 2013). Johansson et al. (2006) measured the effect of internal images by using eye tracking methodology in two scenarios. In the first scenario, viewers inspected a complex picture and afterwards recalled the picture by orally describing it while looking at a blank screen. In the second scenario, viewers looked at a blank screen while listening to a verbal scene description and recalled the scene description by orally retelling it. Their eye movements were measured during both encoding (viewing or listening) and recall (retelling). Results from this study revealed that participants to a very high degree executed eye movements to appropriate spatial locations while describing the picture from memory, while listening to the spoken scene description (that was never seen in the first place) and while retelling it from memory. The effect was equally strong during recall, irrespective of whether the original elicitation was visual or spoken. Here is an obvious link to audio description. By using verbal descriptions, we can evoke and stimulate the creation of vivid internal images. The question is, however, how it works for BVI audiences.

Contemporary research suggests that even people who are congenitally blind experience a kind of mental images, but that their experience differs from that of sighted people in several respects (Cattaneo & Vecchi, 2011; see also Johansson, 2016). Their mental imagery is

often represented in a very spatial manner and instead of being visual, it is more dependent on embodiment and on haptic and motor imagery (e.g., Cattaneo & Vecchi, 2011; Noordzij et al., 2007).

Concerning *information processing*, sighted people have the possibility to process large pieces of information at once (Cornoldi et al., 1989) and use a number of strategies in order to grasp larger entities, forms and global structures in a visuo-spatial environment (Kozhevnikov et al., 2005). Blind people, however, do not have the same kind of overview. To compensate for these limitations, it is common that blind persons complement their mental images with abstract semantic knowledge, which makes their total experience become richer and more effective (Röder & Rösler, 1998). A sequential approach to information processing is evident in the blind persons' preference to take a first-person perspective when they move from one place to another (Noordzij et al., 2007; Postma et al., 2006). This navigation strategy is based on identifying how important landmarks along the way are positioned in relation to one's own body (for example, left or right) and to tie them together into a meaningful unit in a piecemeal way. This is crucial to consider in communication between the sighted and the blind and fundamental for an audio describer to know when selecting what to describe and how (Vandaele, 2007, 2012).

Finally, an area of importance for the understanding and enjoyment of film that has lately gained more scientific interest is *vocal delivery*. Current research shows that the quality of the speaker's voice affects the listening effort no less than listener's attitudes and comprehension of the spoken message (Lyberg-Åhlander et al., 2015; Rogerson & Dodd, 2005). Listeners process the message more slowly and frequently miss content bearing words when listening to a dysphonic (hoarse) voice. This often has direct effects on their comprehension, but it also increases cognitive load and listening effort. Temporal aspects of spoken messages, such as speech rate, fluency and pause distribution are known to be highly important for the understanding of a spoken message and are inherently linked to how verbal information units are segmented in spoken narratives (Eklund, 2004).

Applied to AD, we need to study how voice quality, speech rate and pausing of the audio describer affect understanding and enjoyment (Fryer, 2010). The effect of speed and intonation on the blind and visually impaired users' understanding and enjoyment was investigated by Cabeza-Cáceres (2013). The author found that when AD is delivered at the speed of 14 characters per second, users' comprehension is comparable to that of sighted viewers, whereas at a faster speed, comprehension decreases. Also, intonation, style and vocal delivery may affect reception (Cabeza-Cáceres, 2013; Walczak & Fryer, 2017). The aural dimension (including characters' vocal expression of emotion, soundtrack, ambient sounds and special effects) was highlighted by Fernández et al. (2015). Further, Fernández-Torné and Matamala (2015) conducted research on end-users' preferences concerning text-to-speech versus human voice, as well as standard versus alternative AD. Most participants accept text-to-speech audio description as an alternative solution to the standard human-voiced audio description. However, natural voices obtain statistically higher scores than synthetic voices and are still the preferred solution. Similar results were shown in a small-scale reception study, where BVI participants evaluated AD versions with speech synthesis and with human voice in various genres (Holsanova et al., 2015).

5. Future directions

Audio description is a very complex and creative activity and cognitive aspects underlying the production and reception of AD are still understudied. More empirical research is needed for

understanding how exactly processes of meaning-making function, from a cognitive and user-oriented point of view. By meaning-making, we mean multimodal textual processing, understanding and enjoyment both in the sighted and the BVI audiences. As stated in section 2, there are three types of processes that we still know little about and that need further scrutiny. One of them is *how the audio describers process and understand the multimodal film* (R1), another one concerns *processes underlying production of AD* (P2). The third area that calls for further exploration and experimentation is *how the AD is received by the target BVI audiences* (R2). Future research must map out in detail audio describers' meaning-making, selection, decision-making and mediating activities. Reception studies are necessary as an empirical basis for formulation of linguistic and communicative strategies that would lead to an effective audio description (Braun, 2007; Fresno et al., 2014; Holsanova et al., 2020). There are so far only few empirical studies where identified strategies and solutions have been evaluated by the BVI audiences and explicitly formulated as recommendations. On the other hand, there are many relevant theoretical accounts, research findings and methods from related areas of research that can be employed when studying perceptual and cognitive processes underlying production and reception of AD (cf. section 3 and 4). In order to achieve an effective AD that fulfils the needs of the BVI audiences, various solutions and strategies must be tested by the end-users and results from tests and empirical studies must be translated into AD guidelines and applied for AD practices. Future research must therefore identify users' needs, preferences, associations and conceptualisations, elicit the difficulties BVI audiences are facing during reception of AD and investigate their understanding and involvement. The knowledge obtained from these studies needs then to be incorporated into AD training in order to develop AD-specific comprehension and production strategies (Braun, 2007).

When investigating these processes, scholars should consider using a *broader variety of methods* (cf. section 3). The traditional offline methods such as questionnaires, surveys and interviews that are used to evaluate general attitudes, viewing habits, problems and preferences, could be to a greater extent complemented by dynamic approaches and online methods. Eye tracking, think-aloud protocols and keystroke logging are, for instance, suitable for monitoring visual perception, attention and meaning-making processes during audio describers' reception of the original (R1), and can also be used for monitoring audio describers' selection and decision-making thought processes during AD production (P2). For a detailed investigation of reception processes and for online tracking of emotions, involvement and cognitive load, skin conductance response and EEG seem to be suitable. Offline methods such as ratings, interviews and focus groups can be used to evaluate comprehension, memory, mental imagery and impressions after a concrete film performance.

Another promising future scenario is to combine *descriptive research* with data-driven *experimental studies* and to test the effect of the AD product on actual BVI audiences (Matamala & Remael, 2015). By applying this approach, the solutions and patterns discovered by corpus linguistic and other descriptive studies can be tested and evaluated from a user perspective. Yet another possible approach is to investigate reception of AD by conducting *hypotheses-driven studies* based on various theoretical accounts. A large number of possible *topics* for evaluation and testing has been formulated in current research, for instance:

- which kind of mental images do blind audiences create? Which linguistic solutions should AD entail to enhance mental imagery of BVI audiences? (Holsanova et al., 2020; Johansson, 2016);
- how do blind audiences receive and understand various AD alternatives? (Braun, 2008; Matamala & Remael, 2015);

- how do audiences receive and understand explicit versus implicit verbalisation? Does explicit verbalisation contribute to reducing the processing load (Vandaele, 2012; Braun, 2008; Hirvonen, 2012)? Which other strategies might help to reduce cognitive load? Does anchoring, higher degree of explicitness of temporal setting or selective repetition and vivid presentation of the relevant information decrease processing load for BVI audiences? (Fresno et al., 2014; Vercauteren, 2014);
- how do end-users evaluate AD that convey information about characters' mental states? (Palmer & Salway, 2015);
- how do BVI audiences receive and understand description of gestures, facial expressions and bodily movements? (Mazur, 2014; Vercauteren & Orero, 2013);
- which kind of information do BVI audiences consider most relevant to their needs and which level of detail is optimal for processing? (Calderazzo, 2010);
- do descriptions based on narratological source text analysis improve the BVI audience's understanding of the story? (Vercauteren, 2014).

Another dimension that has not been fully explored is the *vocal delivery of AD* and its role for comprehension and involvement (Braun, 2008; Fryer, 2010). Future studies should, for instance, systematically test how the quality of speaker's voice affects comprehension and listening effort, whether prosody plays a significant role in fore- and backgrounding information, and how speech rate, intonation and style of vocal delivery affect reception of AD (Holsanova et al., 2020).

In future projects, researchers should consider combining *basic and applied research* in their projects. Through basic research, we can gain a better understanding of the principles that underlie successful communication between the sighted and the blind in AD and can systematically study what similarities and differences exist between the way sighted and BVI persons perceive, imagine and understand the contents. These results can then be applied in AD practices and ultimately facilitate the understanding and accessibility of visual information for the BVI audiences (Holsanova et al., 2020).

An interesting issue that has recently been addressed is how AD can be adapted for audiences with cognitive needs, experiencing emotion recognition difficulties (Starr, 2017). Another promising topic that opens up new future opportunities is *computer-generated video description*. There is a need to identify strategies of human-generated descriptions that can inform these automated approaches (Starr et al., 2020; Braun et al., 2020).

Many scholars point out that the field needs an *interdisciplinary approach* (Braun, 2008; Fresno, 2014; Fresno et al., 2014; Holsanova, 2012, 2016b; Reviers, 2017). They highlight the potential of empirical interdisciplinary research combining translation studies, media studies and cognitive studies in order to establish how users process and comprehend audio described films and how they create mental images. It has already been shown that AD can benefit from research methods used in cognitive psychology for exploring narrative delivery and comprehension in AD audiences (Calderazzo, 2010; Kruger, 2010; Fresno, 2014). When we accommodate a wide range of theoretical approaches to AD, integrate insights from various disciplines and use of a combination of quantitative and qualitative methods, we can gain deeper insights into cognitive processes underlying the production and reception of AD (Holsanova, 2016b).

To conclude, research on cognitive aspects of AD concerns the question of how we think, how we perceive information through different senses, how we translate between different senses, how we formulate visual content linguistically and how we represent things and events and create inner images. Knowledge about cognitive processes underlying AD can enrich

models of cognition and communication. A systematic investigation into the dynamics of meaning-making, selection and decision-making processes during production of AD can lead to effective AD strategies and reduce the heavy burden on audio describers' mental resources. Reception research on AD can give valuable insights into how BVI audiences understand, experience and visualise AD descriptions. Last and most important, research on cognitive aspects of AD can enhances AD quality and lead to formulation of guidelines for optimised AD that in turn will be of benefit for the end-users and will facilitate their understanding and immersion.

Notes

1 Meaning-making in a complex multimodal text is an interactive meeting between the recipient, the multimodal message and the situational context (Holsanova, 2014b). All of these three aspects modulate the process of meaning-making. First, the complex multimodal text serves as a starting point for the process of meaning-making through an interplay of various expression modalities. Second, the recipients play an active role and co-create its meaning. Their personal characteristics modulate perception and interpretation of the complex multimodal text. Inter-individual differences arise thanks to the variety of the recipients' backgrounds, interests, previous knowledge, expectations, domain and genre knowledge, expertise, emotions and attitudes. Third, the context in which the visual and verbal parts of the complex multimodal text are displayed, perceived and interpreted plays an important role for meaning-making. In this chapter, by meaning-making, we mean multimodal textual and contextual processing, understanding and enjoyment both in sighted and BVI audiences.
2 These results can be compared to the distinction between descriptive and narrative audio description introduced by Kruger (2010).
3 Vilaró et al. (2012) confirm that not only visual stimuli guide viewers' attention but so too do audio stimuli. The authors show how the soundtrack of audiovisual texts influences perception and comprehension of the scene and highlight that the influence of sound needs to be taken into consideration when producing audio description.

6. Further reading

Braun, S. (2007). Audio description from a discourse perspective: A socially relevant framework for research and training. *Linguistica Antverpiensia NS6*, 357–369.

Holsanova, J. Johansson, R. & Lyberg-Åhlander, V. (2020). How the blind audiences receive and experience audio descriptions of visual events – a project presentation. In *Book of extended abstracts. 3rd Swiss conference on barrier-free communication*. 39–41.

7. References

Andersson, B., Dahl, J., Holmqvist, K., Holsanova, J., Johansson, J.V., Karlsson, H., Strömqvist, H.S., Tufvesson, S. & Wengelin, Å. (2006). Combining keystroke logging with eye tracking. In van Waes, L., Leijten, M. & Neuwirth, C. (Eds.), *Writing and digital media*. Amsterdam: Elsevier. 166–172.

Benecke, B. (2004). Audio description. *Meta: journal des traducteurs*, 49(1), 78-80.

Braun, S. (2007). Audiodescription from a discourse perspective: A socially relevant framework for research and training. *Linguistica Antverpiensia NS6*, 357–369.

Braun, S. (2008). Audio description research: State of the art and beyond. *Translation Studies in the New Millenium*, *6*, 14–30.

Braun, S., Starr, K. & Laaksonen, J. (2020). Comparing human and automated approaches to visual storytelling. In Braun, S. & Starr, K. (Eds.), *Innovations in audio description research*. London: Routledge. 1–12.

Boltz, M. (1992). Temporal accent structure and the remembering of filmed narratives. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 90–105.

Bucher, H.J. (2017). Understanding multimodal meaning-making: Theories of multimodality in the light of reception studies. In Seizov, O. & Wildfeuer, J. (Eds.), *New studies in multimodality: Conceptual and methodological elaborations*. London and New York: Bloomsbury. 91–123.

Cabeza-Cáceres, C. (2013). *Audiodescripció i recepción. Efecte de la velocitat de narració, l'entonació i l'explicitació en la comprensió fílmica*. Unpublished doctoral dissertation, Universitat Autònoma de Barcelona, Spain.

Calderazzo, D. (2010). The "stage in the head": A cognitive approach to understanding audio description in the theatre. *Theatre Topics*, *20*(2), 171–180.

Cámara, L. & Espasa, E. (2011).The audio description of scientific multimedia. *The Translator, 17*(2), 415–437.

Cattaneo, Z. & Vecchi, T. (2011). *Blind vision. The neuroscience of visual impairment*. Cambridge: MIT Press.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*, 293–332.

Chmiel, A. & Mazur, I. (2012). AD reception research: Some methodological considerations. In Perego, E. (Ed.), *Emerging topics in translation: Audio description*. Trieste: EUT. 57–80.

Cornoldi, C., De Beni, R., Roncari, S. & Romano, S. (1989). The effects of imagery instructions on totally congenitally blind recall. *European Journal of Cognitive Psychology*, *1*, 321–331.

Di Giovanni, E. (2014). Visual and narrative priorities of the blind and non-blind: Eye tracking and audio description. *Perspectives: Studies in Translatology*, *22*(1), 136–153.

Eklund, R. (2004). *Disfluency in Swedish human – human and human – machine travel booking dialogues*. Doctoral dissertation, Linköping University Electronic Press.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Fernández, E.I., Martínez, S.M. & Núñez, A.J.C. (2015). Cross-fertilization between reception studies in audio description and interpreting quality assessment: The role of the describer's voice. In Piñero, R.B. & Cintas, J.D. (Eds.), *Audiovisual translation in a global context*. Palgrave studies in translating and interpreting. London: Palgrave Macmillan. 72–95.

Fernández-Torné, A. & Matamala, A. (2015). Text-to-speech vs. human voiced audio descriptions: A reception study in films dubbed into Catalan. *The Journal of Specialised Translation*, *24*, 61–88.

Finke, R.A. (1989). *Principles of Mental Imagery*. Cambridge, MA: MIT Press.

Forceville, C. (2014). Relevance theory as a model for multimodal communication. In Machin, D. (Ed.), *Visual communication*. Berlin: De Gruyter Mouton. 51–70.

Fresno, N. (2014). Is a picture worth a thousand words? The role of memory in audio description. *Across Languages and Cultures*, *15*(1), 111–129.

Fresno, N., Castellà, J. & Soler Vilageliu, O. (2014). Less is more. Effects of the amount of information and its presentation in the recall and reception of audio described characters. *International Journal of Sciences: Basic and Applied Research*, *14*(2), 169–196.

Fryer, L. (2010). Audio description as audio drama – A practitioner's point of view. *Perspectives: Studies in Translatology*, *18*(3), 205–213.

Grice, P. (1975). Logic and conversation. In Cole, P. & Morgan, J.L. (Red.), *Syntax and semantics*, *Vol. 3: Speech acts* (s. 41–58). New York, NY: Academic Press.

Henderson, J. & Ferreira, F. (2004). *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Henderson, J.M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219–222

Hirvonen, M. (2012). Contrasting visual and verbal cueing of space: Strategies and devices in the audio description of film. *New Voices in Translation Studies*, *8*, 21–43.

Hirvonen, M. (2014). *Multimodal representation and intermodal similarity cues of space in the audio description of film*. Academic dissertation, University of Helsinki, Finland.

Holsanova, J. (2001). *Picture viewing and picture description: Two windows on the mind*. Doctoral dissertation, Lund University Cognitive Studies 83.

Holsanova, J. (2008). *Discourse, vision, and cognition*. Amsterdam and Philadelphia: Benjamins.

Holsanova, J. (2011). How we focus attention in picture viewing, picture description, and during mental imagery. In Sachs-Hombach, K. & Totzke, R. (Eds.), *Bilder, Sehen, Denken*. Herbert von Halem Verlag: Köln. 291–313.

Holsanova, J. (2012). Methodologies for multimodal research. *Visual Communication*, Special issue, *11*(3), 251–257.

Holsanova, J. (2014a). Reception of multimodality: Applying eye tracking methodology in multimodal research. In Jewitt, C. (Ed.), *Routledge handbook of multimodal analysis* (2nd edition). London: Routledge. 285–296.

Holsanova, J. (2014b). In the eye of the beholder: Visual communication from a recipient perspective. In Machin, D. (Ed.), *Visual communication*. *Handbooks of communication science [HoCS]*. Berlin/ Boston: De Gruyter. 331–335.

Holsanova, J. (2016a). Cognitive approach to audio description. In: Matamala, A. & Orero, P. (Eds.), *Researching audio description: New approaches*. London: Palgrave Macmillan. 49–73.

Holsanova, J. (2016b). Kognitiva och kommunikativa aspekter av syntolkning. In Holsanova J., Wadensjö, C. & Andrén, M. (Eds.), *Syntolkning – forskning och praktik*. (Audio description – research and practices) Lund: Lund University Cognitive Studies, 166./ Stockholm: Myndigheten för tillgängliga medier, rapport nr. 4. 17–27.

Holsanova, J. (2020a). Att beskriva det som syns men inte hörs: Om syntolkning. To describe what is visible but not audible. On audio description]. *Humanetten*, *44*, 125–146.

Holsanova, J. (2020b). Uncovering scientific and multimodal literacy through audio description. *Journal of Visual Literacy*, *39*(3–4), 132–148.

Holsanova, J., Hildén, A., Salmson, M. & Kesen Tundell, V. (2015). *Audio description and audio subtitles. A study of user preferences with guidelines for audiovisual media*. Stockholm: Tundell & Salmson.

Holsanova, J., Holmberg, N. & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*, *23*(9), 1215–1226.

Holsanova, J. Johansson, R. & Lyberg-Åhlander, V. (2020). How the blind audiences receive and experience audio descriptions of visual events – a project presentation. *Book of Extended Abstracts. 3rd Swiss Conference on Barrier-free Communication*, 39–41.

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.

Jakobson, R. (1959). On linguistic aspects of translation. In Brower, R. (Ed.), *On translation*. Cambridge, MA: Harvard University Press. 232–239.

Jankowska, A. (2019). *AD decision making process. A look inside the describer's head*. Presentation at Media for all 8 conference, Stockholm 2019.

Johansson, R. (2016). Mentala bilder hos seende och blinda. In Holsanova, J., Wadensjö, C. & Andrén, M. (Eds.), *Syntolkning – forskning och praktik*. (Audio description – research and practices) Lund: Lund University Cognitive Studies /Stockholm: MTM:s rapportserie. 29–38.

Johansson, R., Holsanova, J. & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, *30*(6), 1053–1079.

Johansson, R., Holsanova, J. & Holmqvist, K. (2013). Using eye movements and spoken discourse as windows to inner space. In Paradis, C., Hudson, J. & Magnusson, U. (Eds.), *Conceptual spaces and the construal of spatial meaning. Empirical evidence from human communication*. Oxford: Oxford University Press. 9–28.

Kluckhohn, K. (2005). Informationsstrukturierung als Kompensationsstrategie – Audiodeskription und Syntax. In Fix, U. (Ed.). *Hörfilm. Bildkompensation durch Sprache*. Berlin: Erich Schmidt Verlag. 49–66.

Kozhevnikov, M., Kosslyn, S. & Shepard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory & Cognition*, *33*(4), 710–726.

Kress, G. & van Leeuwen, T. (1996/2006). *Reading images: The grammar of visual design*. London: Routledge.

Kress, G. & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Arnold Publishers.

Kruger, J.L. (2010). Audio narration: Re-narrativising film. *Perspectives: Studies in Translatology*, *18*(3), 231–249. doi:10.1080/0907676X.2010.48568.

Kruger, J.L. (2012). Making meaning in AVT: Eye tracking and viewer construction of Narrative. In Mazur, I. & Jan-Louis Kruger, J.L. (Eds.), *Perspectives: Studies in Translatology*, Special Issue. *20*(1), 67–86.

Kruger, J.L., Doherty, S. & Ibrahim, R. (2017). *Beta coherence as objective measure of immersion in film*. Presentation at ARSAD conference Barcelona 2017.

Laeng, B., Bloem, I.M., D'Ascenzo, S. & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition, 131*, 263–283.

Loschky, L.C., Larson, A.M., Magliano, J.P. & Smith, T.J. (2015). What would Jaws do? The tyranny of film and the relationship between Gaze and higher-level narrative film comprehension. *PLoS ONE*, *10*(11), e0142474.

Lyberg-Åhlander, V., Brännström, K.J. & Sahlen, B. (2015). On the interaction of speakers' voice quality, ambient noise and task complexity with children's listening comprehension and cognition. *Frontiers in Psychology*, *6*, 871.

Matamala, A. & Remael, A. (2015). Audio-description reloaded: An analysis of visual scenes in 2012 and Hero. *Translation Studies*, *8*(1), 63–81.

Mayer, R.E. (Ed.) (2005). *Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mazur, I. (2014). Gestures and facial expressions in audio description. In Maszerowska, A. Matamala, A. & Orero, P. (Eds.), A*udio description: New perspectives illustrated*. Amsterdam: Benjamins Publishing Company. 179–197.

Noordzij, M.L., Zuidhoek, S. & Postma, A. (2007). The influence of visual experience on visual and spatial imagery. *Perception*, *36*, 101–112.

Palmer, A. & Salway, A. (2015). Audio description on the thought-action continuum. *Style*, *49*(2), 126–148.

Postma, A., Zuidhoek, S., Kappers, A.M. & Noordzij, M.L. (2006). Haptic spatial orientation pro- cessing and working memory. *Cognitive Processing*, *7*(1), 181.

Radvansky, G.A. & Zacks, J.M. (2014). *Event cognition*. Oxford University Press.

Remael, A., Reviers, N. & Vercauteren, G. (2014). Pictures painted in words. In *ADLAB audio description guidelines*. Trieste: EUT Edizioni Università DiTrieste.

Remael, A., Reviers, N. & Vandekerckhove, R. (2016). From translation studies and audiovisual translation to media accessibility. Some research trends. *Target*, *28*(2), 248–260. doi:10.1075/target.28.2.06rem

Reviers, N. (2017). *Audio description in Dutch: A corpus-based study into the linguistic features of a new, multimodal text type*. Doctoral dissertation, Antwerpen.

Reviers, N. & Remael, A. (2015). Recreating multimodal cohesion in audio description: A case study of audio subtitling in dutch multilingual films. *New Voices in Translation Studies*, *13*, 50–78.

Richardson, D.C., Altmann, G.T.M., Spivey, M.J. & Hoover, M.A. (2009). Much ado about eye movements to nothing: A response to Ferreira et al.: Taking a new look at looking at nothing. *Trends in Cognitive Science, 13*(6), 235–236.

Röder, B. & Rösler, F. (1998). Visual input does not facilitate the scanning of spatial images. *Journal of Mental Imagery*, *22*, 165–181.

Rogerson, J. & Dodd, B. (2005). Is there an effect of dysphonic teachers' voices on children's processing of spoken language? *Journal of Voice*, *19*(1), 47–60.

Smith, T.J. (2013). Watching you watch movies: Using eye tracking to inform cognitive film theory. In Shimamura, A.P. (Ed.), *Psychocinematics: Exploring cognition at the movies*. New York: Oxford University Press. 165–191.

Smith, T.J. & Henderson, J. (2008). Attentional synchrony in static and dynamic scenes. *Journal of Vision*, *8*(6), 773.

Starr, K.L. (2017). *Audio description and cognitive diversity*: *A bespoke approach to facilitating access to the emotional content in multimodal narrative texts for autistic audiences*. PhD. thesis, Surrey University.

Starr, K.L., Braun, S. & Delfani, J. (2020). Taking a cue from the human: Linguistic and visual prompts for the automatic sequencing of multimodal narrative. *Journal of Audiovisual Translation*, *3*(2), 140–169.

Vandaele, J. (2007). What meets the eye. Cognitive narratology for audio description. *Perspectives: Studies in Translatology*, *20*(1), 87–102.

Vandaele, J. (2012). What meets the eye. Cognitive narratology for audio description. In Mazur, I. & Kruger, J.L. (Eds.), *Perspectives: Studies in Translatology*, Special Issue, *20*(1), 87–102.

van Gogh, T. & Scheiter, K. (2009). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*, 95–99.

van Someren, M., Barnard, Y. & Sandberg, J. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press.

Vercauteren, G. (2007). Towards a European guideline for audio-description. In Díaz-Cintas, J., Orero, P. & Remael, A. (Eds.), *Media for all: Subtitling for the deaf, audio description, and sign language*. Amsterdam: Rodopi, 139–150.

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*, *27*(2), 71–90.

Vercauteren, G. (2016). A translational and narratological approach to audio describing narrative characters. *TTR: études sur le texte et ses transformations*, *27*(2), 71–90.

Vercauteren, G. & Orero, P. (2013). Describing facial expressions: Much more than meets the eye. *Quaderns*, *20*, 187–199.

Vercauteren, G. & Remael, A. (2014). Spatio-temporal settings. In Maszerowska, A., Matamala, A. & Orero, P. (Eds.), *Audio description: New perspectives illustrated*. Amsterdam and Philadelphia: John Benjamins. 61–80.

Vilaró, A.,Duchowski, A., Orero, P., Grindinger, T., Tetreault, S. & di Giovanni, E. (2012): How sound is The pear tree story? Testing the effect of varying audio stimuli on visual attention distribution. *Perspectives: Studies in Translatology,* 20(1): 55–65.

Walczak, A. & Fryer, L. (2017). Vocal delivery of audio description by genre: Measuring users' presence. *Perspectives Studies in Translatology, 26*(1), 1–15.

Wildfeuer, J. (2012). More than words. Semantic continuity in moving images. *Image & Narrative*, *13*(4), 181–203.

Wilson, D. & Sperber, D. (2004). Relevance theory. In Horn, L. & Ward, G. (Eds.), *The handbook of pragmatics*. Oxford: Blackwell Publishing. 607–632.

Zabalbeascoa, P. (2008). The nature of the audiovisual text and its parameters. In Díaz Cintas, J. (Ed.), *The didactics of audiovisual translation*. Amsterdam: John Benjamins Publishing Company. 21–37.

Zacks, J.M., Speer, N.K. & Reynolds, J.R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology*: *General*, *138*(2), 307–327.

Zwaan, R.A., Magliano, J.P. & Graesser, A.C. (1995). Dimensions of situation-model construc- tion in narrative comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *21*, 386–397.

Zwaan, R.A. & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.

8. Acknowledgements

This work was supported by a grant from FORTE 2018–00200 (Swedish Research Council for Health, Working Life and Welfare).