LUND UNIVERSITY

Event segmentation in the audio description of films

A case study

Holsanova, Jana; Blomberg, Johan; Blomberg, Frida; Gärdenfors, Peter; Johansson, Roger

Published in:
Journal of Audiovisual Translation

DOI:
10.47476/jat.v6i1.2023.245

2023

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Holsanova, J., Blomberg, J., Blomberg, F., Gärdenfors, P., & Johansson, R. (2023). Event segmentation in the audio description of films: A case study. *Journal of Audiovisual Translation*, *6*(1), 64-92. https://doi.org/10.47476/jat.v6i1.2023.245

Total number of authors:
5

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Journal of Audiovisual Translation

Event Segmentation in the Audio Description of Films: A Case Study

Jana Holsanova ✉
Cognitive Science Department, Lund university, Sweden

Johan Blomberg ✉
Department of Philosophy, Linguistics and Theory of Science, Gothenburg university, Centre for Languages and Literature, Lund university, Sweden

Frida Blomberg ✉
Academic Support Centre, Lund university, Sweden

Peter Gärdenfors ✉
Cognitive Science Department, Lund university, Sweden,
Paleo-Research Institute, University of Johannesburg, South Africa

Roger Johansson ✉
Psychology Department, Lund university, Sweden

Citation: Holsanova, J., Blomberg, J., Blomberg, F., Gärdenfors, P., & Johansson, R. (2023). Event Segmentation in the Audio Description of Films: A Case Study. *Journal of Audiovisual Translation, 6*(1), 64–92. https://doi.org/0.47476/jat.v6i1.2023.245

Editor(s): N. Reviers

Received: June 13, 2022

Accepted: May 10, 2023

Published: December 7, 2023

Copyright: ©2023 Holsanova, J., Blomberg, J., Blomberg, F., Gärdenfors, P., & Johansson, R. This is an open access article distributed under the terms of the Creative Commons Attribution License. This allows for unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Acknowledgments

This work was supported by a grant from FORTE 2018–00200 (Swedish Research Council for Health, Working Life and Welfare) and by TSI 2019 Lund University grant "Audio Description and Accessible Information" (2021–2025).

Abstract

To make the content of films available to a visually impaired audience, a sighted translator can provide audio description (AD), a verbal description of visual events. To achieve this goal, the audio describer needs to select *what* to describe, *when* to describe it, and *how* to describe it, as well as to express the information aurally. The efficacy of this communication is critically dependent upon basic cognitive processes of how the sighted audio describer perceives and segments the film's unfolding chain of events and in what way the visually impaired end users conceive the structure, content, and segmentation of such events in relation to the produced AD. There is, however,

✉ jana.holsanova@lucs.lu.se, https://orcid.org/0000-0002-7510-6526
✉ johan.blomberg@semiotik.lu.se, https://orcid.org/ 0000-0002-1337-1981
✉ blomberg.frida.m@gmail.com
✉ peter.gardenfors@lucs.lu.se, https://orcid.org/ 0000-0001-7423-828X
✉ roger.johansson@psy.lu.se, https://orcid.org/ 0000-0003-3434-2538

virtually no research on this interplay in relation to AD. In this study, we scrutinize live AD of a film from two trained audio describers and examine how events are structured, segmented and construed in their AD. Results demonstrate that the event segmentation experienced from the film is indeed a fundamental part of how AD is structured and construed. It was found that AD at event boundaries was highly sensitive to different spatiotemporal circumstances and this relationship depends on semantic resources for expressing AD.

Key words: audio description, event segmentation, event cognition, film, language production, event semantics.

Introduction

To make the content of moving images and audio-visual media available to a visually impaired audience, a sighted translator can provide audio description (AD). AD comprises primarily verbal descriptions of visual events, aiming to increase the accessibility of visual information and to provide a visually impaired audience with a richer and more detailed understanding, experience, and enjoyment of, for instance, films and TV shows. To achieve this, the audio describer selects relevant information from the visual event (e.g., environments, objects, people and their appearance, clothing, facial expressions, actions, gestures, and body movements) and expresses this information aurally, by using vivid verbal descriptions. In this way, the audio describer can evoke inner mental images for the visually impaired audiences and enhance their meaning-making (Holsanova, 2016; 2022; Holsanova et al., 2016; Johansson 2016; Vandaele, 2012). When producing AD, audio describers must also integrate their verbal descriptions with sounds, voices, and dialogues from films, putting high demands on the timing of the narration process. For instance, they should not talk during dialogues, and must judge when to explain certain sounds (e.g., who is shooting whom, when shots are heard) and when to let them "speak" for themselves (e.g., when there's a knocking on the door, when the phone rings, when everybody is laughing). This task is particularly challenging when AD is produced online in a live setting. Other challenges include the richness of the visual mode, the need to prioritise information and to select appropriate verbal means for describing the visual content. In essence, when producing AD, the audio describer constantly needs to evaluate *what* to describe, *when* to describe it, and *how* to describe it (Holsanova, 2016; 2022; Holsanova et al., 2016; Vercauteren, 2021). The efficacy of this communication is critically dependent upon basic cognitive processes of how the sighted audio describer perceives and segments the film's unfolding *chain of events* and in what way the visually impaired end users conceive the structure, content, and segmentation of such events in relation to the produced AD and the sound expressions from the film (cf. Vandaele, 2012).

The way in which events are structured and distinguished from one another has been the matter of much research over the last three decades and there is extensive evidence that we, as cognitive human beings, are equipped with automatic neurocognitive principles to perceive events as distinct from one another (for overviews, see Radvansky & Zacks, 2014; Kurby & Zacks, 2008; Zwaan & Radvansky, 1998). Thus, a fundamental question in successful AD should concern how the sequence of events unfolding in the movie relates to the verbal narration communicated through AD. To our knowledge, no previous studies have investigated event segmentation in AD (but see Vercauteren, 2021, where concepts from event cognition and mental model theory are used to analyse content selection in AD).

Event segmentation is also important for computer-generated video description, a new promising area that is being rapidly developed to supplement human audio description and make audio-visual media more widely accessible. However, as the comparisons of human and automatic scene

descriptions show, there are fundamental problems with the current state and quality of machine-generated descriptions (Braun et al., 2020; Starr et al., 2020). Among other issues, computer algorithms "see" images in isolation as single frames, do not integrate them, and are likely to miss several key actions (or mis-label them). An even more pressing issue is that current algorithms are blind to narrative cohesion across individual frames, shots, and scenes. The quality of automated video description could be improved if the training of algorithms could focus on identifying actions and events in dynamic scenes, and on recognising event boundaries.

In the present case study, we scrutinize how events are structured, segmented, and construed in AD for a commercial film. As we utilise methods to analyse spoken discourse that have been developed in studies on spontaneous discourse, we have chosen to focus on live AD, which more closely corresponds to such circumstances. In live AD, the audio describer is physically present in the movie theatre and produces AD online. The audience listen to the AD on their headphones and at the same time hear the original sound played from the ordinary sound system in the movie theatre. However, while the AD is produced live, the audio describer is well prepared and has seen the film before. Considering that ADs of films are typically pre-recorded and applied offline by audio describers who are using a more "prepared" discourse, the present case study is somewhat atypical. Nonetheless, live AD is still an important and integral part of the AD culture and is highly appreciated by audiences. For instance, in Sweden, live AD is commonly offered during film festivals[1] and when certain associations arrange their own screenings and special events.

We analysed two ADs of the same commercial film, produced live by two different professional audio describers.

1. Event Segmentation

Our everyday life unfolds over time, with enormous amounts of multifaceted information constantly getting intertwined into a continuous stream of experience. However, the mental models that we as cognitive human beings construct of the world are not continuous. Instead, like scenes in a movie, we understand and remember our experiences as distinct and meaningful chains of events (cf. Radvansky & Zacks, 2014). Such events can be defined as "a segment of time at a given location that is conceived by an observer to have a beginning and an end" (Zacks & Tversky, 2001; Kurby & Zacks, 2008). An accumulating body of research has demonstrated that there are extensive overlaps across people in what they perceive as distinct meaningful events (e.g., Newtson, 1973; Huff et al., 2014; Kurby & Zacks, 2011; Speer et al., 2003; Zacks & Tversky, 2001; Zacks et al., 2009), suggesting that there are generic cognitive principles for how we "automatically" process and segment event

[1] At *Gothenburgs* film festival 2023, 6 selected films were screened with live AD. At *Stockholm* film festival 2021, 6 selected films were screened with live AD. At the upcoming Children and Youth Film Festival in Malmoe 2023, a few films will be screened with live AD.

information. For instance, there is considerable agreement among participants when they are asked to subjectively indicate what they experience as meaningful event units in narrative texts and films (e.g., Newtson, 1973; Zacks et al., 2009; Zacks & Tversky, 2001). Corroborating evidence comes from current neurocognitive research, where the neural underpinnings that subserve event segmentation have been mapped out (e.g., Amoruso et al., 2013; Baldasssano et al., 2018; Ezzyat & Davachi, 2011; Zacks et al., 2010; Stawarczyk et al., 2020; Zacks et al., 2001). For instance, transient changes in specific brain activity have been demonstrated to occur similarly during passive viewing of films and during active event segmentation (Zacks et al., 2001).

A narrative, such as a film or a verbally told story, is already segmented into a chain of events coordinated in different ways. Sometimes the chain is not presented in temporal order but is broken up, for example including a flashback or flashforward. To make sense of narrative coherence – or lack thereof – one must be able to keep track of how different events are related to one another. In contrast to "natural" event segmentation, a film typically also contains cuts that create discontinuities and unorthodox corresponding event boundaries differing from the natural sequence of experience (e.g., Zacks, 2013).

In this section, we outline models for segmenting events in both visual cognition and language production. With respect to the former, Zacks et al. (2009) and Cutting and Iricinschi (2015) propose accounts of event segmentation applied to films. They suggest that there are different kinds of indicators signalling beginnings and ends of events, such as changes in time, place, and characters. We outline and evaluate these accounts before proposing three types of event changes that we deem to be of specific relevance for the present study.

How language structures events and how "event units" are related to one another are big topics in cognitive linguistics (e.g., Pustejovsky, 1991; Talmy, 2000); it has been proposed that the semantic resources for expressing events can be generalized to comprise *how* an action is carried out – the *manner* – and *what* the action leads to – the *result* (e.g., Gärdenfors, 2014). While linguistic event semantics have traditionally not been concerned with perception of non-verbal event segmentation, recent evidence suggests that the perception of visually presented stimuli and spoken event descriptions of corresponding information rely, to a large degree, on the same event segmentation principles (Gerwien & von Stutterheim, 2018).

1.1. Event Segmentation, Films and AD

As experiences can be broken down into sequences of events, this means that there are ways to categorize and distinguish events. Even though it might be quite clear that there are boundaries between events, exactly when and how such boundaries are construed is less discernible. Thus, to investigate event segmentation, qualitative aspects that mark the transition from one event to

another need to be determined. Several different proposals on how to identify event boundaries have been made, primarily based on studies of narratives in different media. One of the more influential models is the *event-indexing model* (Zwaan et al., 1995; Zwaan & Radvansky, 1998). Based on an analysis of written narratives in literature (fiction), the model proposes several indices involved in processing of a narrative, where relationships between characters, time and space are fundamental for event segmentation. In making sense of a narrative, one is required to keep track of such indices and distinguish between chains of events that take place at different locations, at different times and/or involve different characters that act in relation to distinct spatiotemporal contexts. Inspired by the event-indexing model, subsequent studies have extended it to investigating films, and two prominent accounts to classify and segment films have been proposed by Zacks et al. (2009) and Cutting and Iricinschi (2015). While their respective accounts are largely compatible – identifying changes in time and space as crucial for segmenting events – there are also considerable differences between the two. Zacks et al. (2009) identify six different types of event changes: Cause, Character, Goal, Object, Space, and Time. They argue that these types of changes are marked in visually narrated media by techniques such as cutting and panning. Cutting & Iricinschi (2015) use only three different categories: Location, Time, and Character change. However, they emphasize the theoretical difficulty in defining them systematically and the practical problem in univocally assigning change type to different events in a film. A film may change between different historical epochs and locations, for instance between Sweden in the 1990s and Cuba in the 1960s. In such an example, there are both temporal and location changes. At other times, the change in location can be somewhat continuous within a scene, such as entering or exiting a room.

In the present study, we focus on changes in time and location. The main reason for this is that temporal and locational changes are salient categories, which both have been in focus in previous studies (e.g., Speer et al., 2003; Speer & Zacks, 2005; Zacks et al., 2010), and importantly, they are expected to be expressed by the audio describer (see Vercauteren, 2021; Rai et al., 2010). Also, spatiotemporal settings belong to the basic narrative building blocks in storytelling and are intrinsically linked to the story's characters and their actions. For instance, spatiotemporal settings can be linked to different characters (e.g., characters that are associated with a particular pub, car, school, or city), or represent different time periods (e.g., flashbacks to a setting in the 60s versus present time). Spatiotemporal settings have been demonstrated to constitute a critical part of AD (Remael & Vercauteren, 2015; Vandaele, 2012; Vercauteren, 2021; Vercauteren & Remael, 2015), and most of the existing guidelines on AD recommend that spatiotemporal features are to be included in the descriptions (cf. Rai et al., 2010; Vercauteren & Remael, 2015). With respect to locational changes, we found it important to make a distinction between those that involved a *change between locations* and a *change within location*. The former involves a location change marking a discrete and distinct change in the spatial context as defined by physical boundaries, such as a change from being inside to being outside of a building. The latter involves changes within one and the same spatial context, such as a change from the living room to the bedroom inside the same apartment. While the difference between those two types of location changes is cognitively salient

(as known from literature on events in linguistics and cognitive science, cf. Talmy, 2000; Gärdenfors, 2014), this distinction has, to our knowledge, not previously been considered in the literature on event segmentation.[2]

Considering temporal changes, Cutting and Iricinschi (2015) note that it can be difficult to discern when there is a shift in time, and that films often involve different characters in successive scenes, where temporal transition is ambiguous. Temporal event boundaries can, for instance, be indicated by "marked shifts in the weather; through flashbacks and flashes forward in time or through shifts in mental state from diegetic reality to dreams or memories and back" (Curreing & Iricinschi, 2015, p. 444). In studies of AD, Vercauteren (2012) has provided extensive details on how temporal settings and narration can be rendered and expressed in AD. For instance, it is common that temporal event boundaries are expressed in a brief and explicit format, such as "at night", "the following morning", "the next day".

We did not consider character change, as such event boundaries are expected to typically be expressed in audio descriptions to a lesser degree, given that the blind audience can in many cases infer character change from the dialogue. Also, character changes are commonly accompanied by changes in the spatiotemporal framework (Cutting & Iricinschi, 2015; Zacks et al., 2009). However, even if descriptions of character change were not considered for event boundaries in the present study, it is important to note that identification and description of characters is an integral and fundamental part of effective AD (see Mazur, 2015; Vercauteren, 2016). For instance, naming of characters and descriptions of their appearance, have been shown to be critical for how characters are perceived and remembered (Benecke, 2014; Fresno, 2012), and how this may influence understanding of the narrative (Vercauteren, 2016).

1.2. Event Segmentation and Spoken Discourse

In cognitive semantics, "events" are inherently embedded in linguistic material, primarily verbs, and typically refer to temporal changes of states, actions, and occurrences (e.g. Klein, 2009; Lakoff, 1987; Pustejovsky, 1991; Talmy, 2000). In the semantic model of events put forward by Warglien et al. (2012), it is proposed that sentences express events, or, more precisely, *construals* of events, suggesting that the linguistic expression of events has an internal semantic structure. In this model, every event can be profiled in different ways, which means that in referring to the same situation, a speaker has some freedom in how the situation is construed. One can, for instance, focus on the

[2] The difference between these types of semantic information forms the basis for an influential cross-linguistic typology of events. Many languages have constraints against combining the expression of Manner and Result in the same clause, e.g., Romance languages. It has been shown that even when grammatically possible, speakers of Romance languages prefer to keep information about an action and its result in separate clauses (Slobin, 1996).

result of an event or the *manner,* in which it was carried out. For instance, if the verb describing a situation denotes the "change vector" (e.g., "move", "walk", "climb", "break") it is a result verb, whereas if it denotes the "force vector" (e.g., "push", "hit", "run") it is a manner verb (Gärdenfors, 2014; Warglien et al., 2012). In cognitive neuroscience, it has been demonstrated that the processing of manner and result verbs activates separable neural systems (Wu et al., 2008), which substantiate the importance of examining these semantic representations in relation to event segmentation. Moreover, memory for events has been demonstrated to be highly sensitive to result and manner information, both when listening to and producing event narratives (Santin et al., 2021; Skordos et al., 2020).

However, in language production, semantic representations of events are not only internally structured, but also follow certain principles for marking the change between discrete events. The ways in which speakers segment discourse and create transitions typically reflect a certain "cognitive rhythm" in discourse production (Holsanova, 2001, 2008). Speech is produced in spurt-like portions and contains disfluencies such as pauses, hesitations, interruptions, restarts, etc. These phenomena provide clues about underlying cognitive processes, such as perception, comprehension, planning, and production.

Most theoretical accounts agree that we are only able to focus on small pieces of information at a time, but there are different ideas concerning the size and form of these information units (Chafe, 1996; Halliday, 1985; Haviland & Clark, 1974; Givón, 1990). Also, there are different traditions regarding what the basic discourse structuring elements are called: idea units, intonation units, information units, information packages, phrasing units, etc. In Chafe's view, the information flow in spoken discourse is associated with dynamic changes in language and thought: an upcoming utterance expresses a change in the information states of both speaker and hearer (Chafe, 1996). This model implies that one new idea is formulated at a time and that focused information is replaced by another idea at short intervals of approximately two seconds (Chafe, 1987). This small unit of discourse (called "idea unit" or "intonation unit") expresses the conscious focus of attention. According to Chafe, such a unit consists of the amount of information to which a person can devote their executive attention at a given point in time. The expression of one new idea at a time reflects a fundamental temporal constraint on the way that the mind can process information (cf. Holsanova, 2001, 2008).

Chafe's model of idea units has been successfully applied and extended in studies of online spoken descriptions of visual scenes (Holsanova, 2001, 2008; Johansson et al., 2013). Here, a *verbal focus* (referring to a specific visual element of the scene) comprises the basic discourse-structuring element, which is usually expressed as a phrase or a short clause, delimited by prosodic, acoustic features and lexical/semantic features. It has one primary accent, a coherent intonation contour and

is usually preceded by a pause, hesitation, or a discourse marker (Holsanova, 2001). Example 1 from a spoken description of a static image illustrates these units (Holsanova, 2008, p. 58).[3]

Example 1

in the middle of the picture there's a **tree'**

uh . and in it three **birds** are sitting

---

in ... to the left in the **tree**,

at the very **bottom'** or or the **lowest** bird on the left in the tree'

she sits on her **eggs'**

uuh and above her'

still on the left'

there is a a a **bigger** bird'

standing up'

Several verbal foci are clustered into a more complex *verbal superfocus*. This larger discourse segment, typically a longer utterance, consists of several foci connected by the same thematic aspect and has a sentence-final prosodic pattern (often a falling intonation). A superfocus is typically preceded by a long pause and a hesitation, reflecting the process of refocusing attention from one scene element to another, but can also be expressed through the aid of specific discourse markers, such as voice loudness, voice quality, or speech tempo. Prosody seems to play a significant role in fore- and backgrounding information. For example, a speech passage uttered in a deviating voice quality (creaky voice or dialect-imitating voice), can stretch over several verbal foci and form a superfocus when describing a particular aspect of a scene. The acoustic features can then act to underscore the beginning and the end of the superfoci borders, and previous studies have shown that listeners easily identify, with high agreement, on unit borders in spoken discourse (Holsanova, 2001, 2008).

In the present study, the principles outlined by Chafe (1987; 1996) and Holsanova (2001, 2008) were used as the basis for segmenting the audio describers' spoken discourse into discrete meaningful

[3] Each line of the transcript represents a verbal focus. A short pause is marked by a point (.), whereas a long pause is marked by three points (…). Rising intonation is marked by an apostrophe ('), falling intonation is marked by a comma (,), and stressing is marked by bold style (**sun**). A superfocus is marked by dashed lines.

intonation units, and the semantic model of events developed by Warglien et al. (2012) and Gärdenfors (2014) was used to code the verbal content into result and manner information.[4]

1.3. Present Study

The central aim of the present case study was to examine how events are structured, segmented and construed in spontaneous discourse generated in live audio descriptions of the Swedish film, *Skumtimmen*. Specifically, we focused on the following research questions:

- How is event information expressed in the rapid real-time spoken language descriptions generated by audio describers?
- To what degree are spatiotemporal event boundaries explicitly verbalised by audio describers?
- How are spatiotemporal event changes verbally construed in relation to manner and result information?

To achieve our goals, we first segmented the film according to spatiotemporal event boundaries based on the event-indexing model (Zwaan et al., 1995; Zwaan et al., 1995; Zwaan & Radvansky, 1998) and its adaptation to motion pictures (Zacks et al., 2010; Cutting & Iricinschi, 2015). Second, we transcribed and coded the two audio descriptions of the film into intonation units (Chafe, 1996; Holsanova, 2001, 2008) and compared them to the identified spatiotemporal event boundaries in the film. Third, we examined the linguistic content of the ADs in relation to the semantic structures of result and manner information (Warglien et al., 2012; Gärdenfors, 2014).

2. Method

2.1. Material

Two ADs of the Swedish film *Skumtimmen* (Alfredson, 2013) were recorded during visually interpreted screenings in a movie theatre. *Skumtimmen* is a Swedish-produced film with a theatrical release in 2013 (English title: *Echoes from the Dead*). The film is a drama-thriller centred around Julia whose son mysteriously disappeared 21 years ago. Now Julia has returned to her childhood home and the truth of past events is starting to resurface. The film takes place over three different time periods and has a runtime of 99 min. The ADs were conducted live by two different Swedish audio

[4] Spoken live AD and spoken online descriptions of visual scenes have several commonalties: they are produced in spurt-like portions, contain disfluencies, and reflect a certain rhythm. However, there are also differences: AD is produced during limited time slots and is not a stand-alone spoken discourse since it functions in combination with sounds, music, and dialogue.

describers on two different occasions. The audio describers had seen the film beforehand but did not create a manuscript. During the live AD, they were required to construct on-the-fly descriptions of the visual information in tandem with the visual stream of the film. The recordings were made with a speech dictation device (Olympus VN 711).

2.2. Coding the Film

The present study focused on spatiotemporal changes and segmented the film based on the event-index model (Zwaan et al., 1995; Zwaan et al., 1995; Zwaan & Radvansky, 1998) and its adaptation to motion pictures (Zacks et al., 2010; Cutting & Iricinschi, 2015). See Figure 1 for an example. The film was segmented into the following three event boundary categories:

- *Change in time (CIT)*: Involves either different points in time (such as past and present, flashbacks, shifts between different periods, e.g., between World War II and the Vietnam War) or time of the day (morning, evening, etc.). Temporal changes are common at scene changes (Cutting & Iricinschi, 2015).
- *Change between locations (CBL)*: Involves a transition from one location setting to another within the same time frame (thus a change in location across different time frames was not coded as CBL, but as CIT), such as from the outside to the inside of a house. Often marked in film by physical boundaries such as doors and gates (Cutting & Iricinschi, 2015).
- *Change within location (CWL)*: In contrast to a change between distinct and different location settings (as in CBL), this category involves a change within one and the same location setting, such as a character moving from one room to another in the same house.

Two independent coders analysed the film *Skumtimmen* according to those categories and determined that it comprised 194 event boundaries in total. Interrater reliability between the coders was high (Cohen's kappa = 0.85, $p < .001$). After their independent categorizations, the coders discussed the deviating categorizations, coming to agreement concerning each. The frequency of the different event boundary categorizations is listed in Table 1. As can be seen, CBL was more common than both CWL and CIT.

Table 1

The proportion of spatiotemporal event boundaries (N = 194) in Skumtimmen

	N	*%*
Change In Time (CIT)	26	13
Change Between Locations (CBL)	104	54
Change Within Locations (CWL)	64	33

2.3. Coding the Spoken Discourse

2.3.1. Identifying Intonation Units

Based on the models and accounts outlined by Chafe (1987; 1994) and Holsanova (2001, 2008), the data was broken down into distinct *intonation units*. A combination of semantic and prosodic criteria was used for identifying and differentiating intonation units (outlined below). The units were identified and coded by two independent coders. See Figure 1 for an example.

Intonation units may vary in length and have quite different semantic and pragmatic purposes. Prosodically, they are, however, characterized by their "coherent intonation contour" with peaks in intonation and cadence that indicate their end. What belongs together thematically is also produced in spoken discourse as belonging together. From the semantic point of view, an intonation unit can be expected to express a single and conceptually coherent event, i.e., a defined number of participants in a determinate spatiotemporal context with a beginning and an end. The simplest example (in Swedish as well as English) would be one agent (expressed by a noun, pronoun, or proper name), which together with a single verb forms the expression of an event. Examples of such minimal event expressions with a Manner-verb (force vector) and Result-verb (change vector) are shown in (1) and (2), respectively.

1. Julia städar.
"Julia is cleaning."

2. Gerloff sitter.
"Gerloff sits."

An event can, and often is, described in much more detail with semantically specific information about both Manner (*småspringer*) and Result (*mot bilen*), see (3).

3. Lennart småspringer mot bilen.
"Lennart is half running towards the car."

With the help of these criteria for defining an intonation unit, we were able to exclude several linguistic strategies from forming new intonation units. Among them are pseudo-coordinations, where two semantically different verbs are conjoined to form a cohesive whole. This is a very common discursive strategy in Swedish (Kvist Darnell, 2008), as exemplified in (4).

4. Julia sitter och tittar.
"Julia sits and watches"

On many occasions, an event was described in one initial intonation unit (the event boundary), and then further information about this event was added in successive intonation units. This often occurred rapidly and was typically characterized by elaborations, additions, and further descriptions

of properties of either the characters or the event (cf. example below)[5]. Thus, whereas an event boundary is always expressed through the initial intonation unit, the whole event may comprise several intonation units (cf. Example 2 below and Figure 1).

Example 2

Här i huset . sitter en kille i **20-årsåldern**'{Event boundary}

*here in the house . sits a guy in his **20s***

… håller på med sin **hagelbössa**,

… working with his shotgun,

han gör den **ren**,

*he **cleans** it,*

han smörjer in den med **olja'**

*he lubricates it with **oil***

putsar på den'

***polishes** it*

… den här unge killen är ute på **Alvaret** nu' {Event boundary}

this young guy is now outdoors on Alvaret

solen skiner'

*the **sun** is shining'*

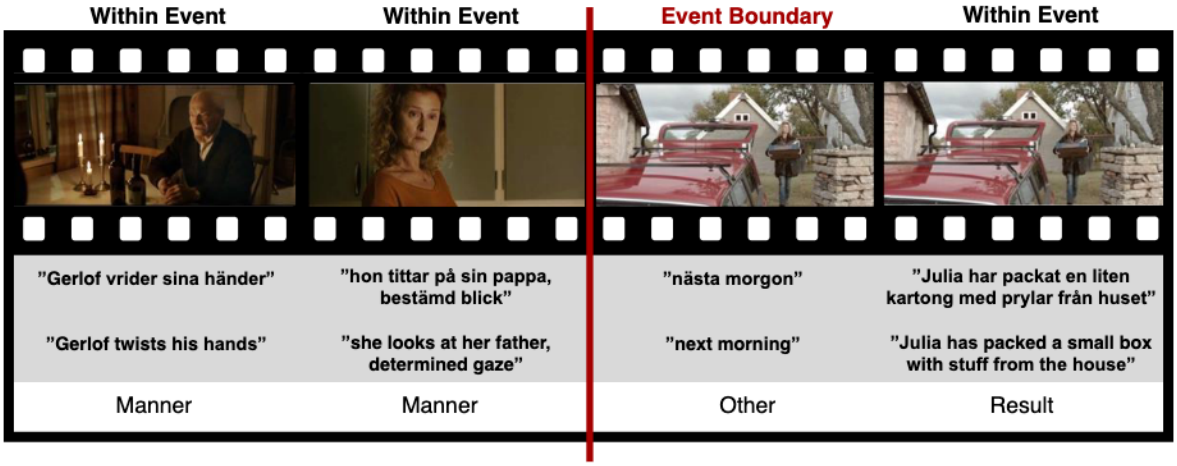
2.3.2. Semantic Coding

Based on semantic structures outlined by Warglien et al. (2012) and Gärdenfors (2014), the two independent coders identified intonation units in ADs produced by the two audio describers and coded them for the expression of Result, Manner, or both (see Section 2.1 and Appendix A). Interrater reliability between the coders was relatively high (Cohen's kappa = 0.63, $p < .001$; to interpret the outcome of this measure, see McHugh, 2012). After their independent categorizations, the coders discussed their deviating categorizations, coming to agreement concerning each. See Figure 1 for an example.

[5] Each line of the transcript represents an intonation unit. A short pause is marked by a point (.), whereas a long pause is marked by three points (…). Rising intonation is marked by an apostrophe ('), falling intonation is marked by a comma (,), and stressing is marked by bold style (**sun**).

Figure 1

Synchronization of intonation units in AD in relation to visual information in the film



Note. Schematics shows the synchronization of intonation units expressed in the AD by one of the audio describers in relation to corresponding visual information in the film. The first two frames illustrate Within Event information (i.e., information given in between two event boundaries). Here the AD provides manner information about two key characters who are having a conversation in the same room. In the third frame, there is an event boundary. Here, the AD describes the temporal event boundary of the next day (without any result or manner information). In the fourth frame, the AD describes result information about one of the characters within this new event. The top panel illustrates visual information present in the film. The centre panel (grey) shows the coded intonation units uttered by the audio describer in relation to this visual information. The bottom panel (white) shows how the verbal content was coded according to the semantic structures of result and manner. The red vertical line illustrates the coded event boundary in the film.

2.4. Analytical Approach

First, we present descriptive statistics of the two audio describers' intonation units in relation to: (1) Event Boundaries versus Within Events (i.e., intonation units corresponding to information in between two event boundaries); (2) the three Event Boundary Categories – change in time (CIT), change between locations (CBL), change within location (CWL); (3) expressions of Manner and Result. Second, we analyse how often the identified event boundaries were explicitly expressed by the two audio describers. Third, we analyse how frequently Manner and Result information was expressed at Event Boundaries versus Within Events. Fourth, we analyse how frequently Manner and Result information was expressed in relation to the three Event Boundary categories. Finally, we exemplify and analyse some characteristic ways of expressing different types of event boundaries in the context of our results.

Statistical analyses are, when applicable, conducted with generalized linear models (Gallucci, 2019) using Jamovi version 1.6.23 (The Jamovi Project, 2019). By considering data from all data points, these models have more power when compared to traditional analyses of variance and allowed us to statistically contrast our different categories of intonation units and semantic expressions over the total data generated by our two audio describers.

3. Results

3.1. Intonation Units and Event Boundaries

In total, the two audio describers generated 1476 intonation units. Of those, about a fifth were intonation units that expressed and corresponded to the identified event boundaries. Within event information, i.e., event information in between event boundaries, was on average expressed with 5.3 intonation units (*SD* = 4.8) for Audio Describer 1 and with 3.5 intonation units (*SD* = 3.4) for Audio Describer 2. Thus, Audio Describer 1 overall produced more verbal content than Audio Describer 2. See Table 2 for descriptive data.

Table 2

Number of uttered intonation units for two audio describers, distributed in relation to event boundary versus within event; over three categories of boundaries

	Total	Audio Describer 1	Audio Describer 2
Intonation units			
Total (N)	1476	920	556
Event boundary (N)	321	166	155
Within event (N)	1155	754	401
Event boundary (%)	22	18	28
Within event (%)	78	82	72
Event boundaries			
CIT (N)	49	26	23
CBL (N)	177	93	84
CWL (N)	95	47	48
CIT (%)	15	16	15
CBL (%)	55	56	54
CWL (%)	30	28	31
Event semantics			
Manner (N)	472	305	167
Result (N)	475	271	204
Both (N)	174	97	77
Other (N)	355	247	108
Manner (%)	32	33	30
Result (%)	32	29	37
Both (%)	12	11	14
Other (%)	24	27	19

Note. Number of uttered intonation units produced by the two audio describers and their distribution in relation to event boundary versus within event information; over the three categories of event boundaries: change in time (CIT), change between locations (CBL), change within location (CWL); and over the semantic categories of Manner, Result, Both or Other.

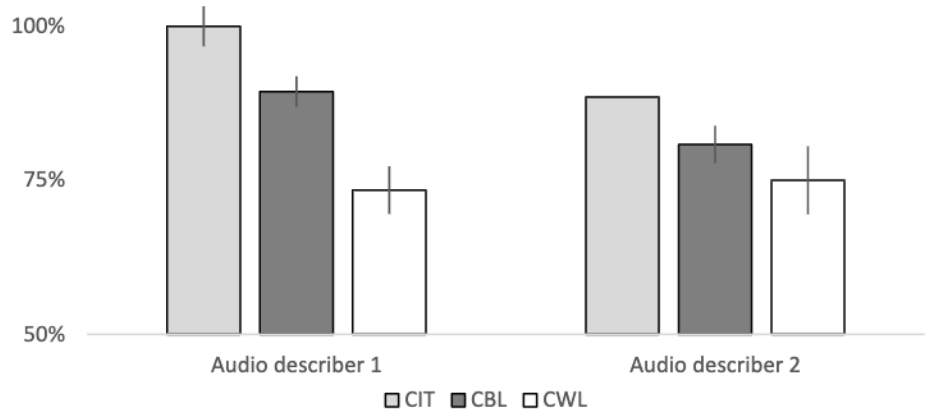
Source: Authors' own study.

Statistical analyses of event boundaries revealed that the category of event boundaries was a significant predictor of whether an event boundary was expressed or not in audio description, $\chi^2(1) = 12.8$, $p = .002$. CWL boundaries were significantly less common to be expressed than both CIT ($p = .006$) and CBL ($p = .015$). No significant difference emerged between CIT and CBL ($p = .093$). Thus,

changes within location were least likely to get explicitly expressed in audio description, whereas changes in time were the most likely. See Figure 2.

Figure 2

Expressed Event Boundaries



Note. Proportion of event boundaries that were explicitly expressed in audio description across the three event boundary categories: change in time (CIT), change between locations (CBL), change within location (CWL). Error bars denote SEM.

Source: Authors' own study.

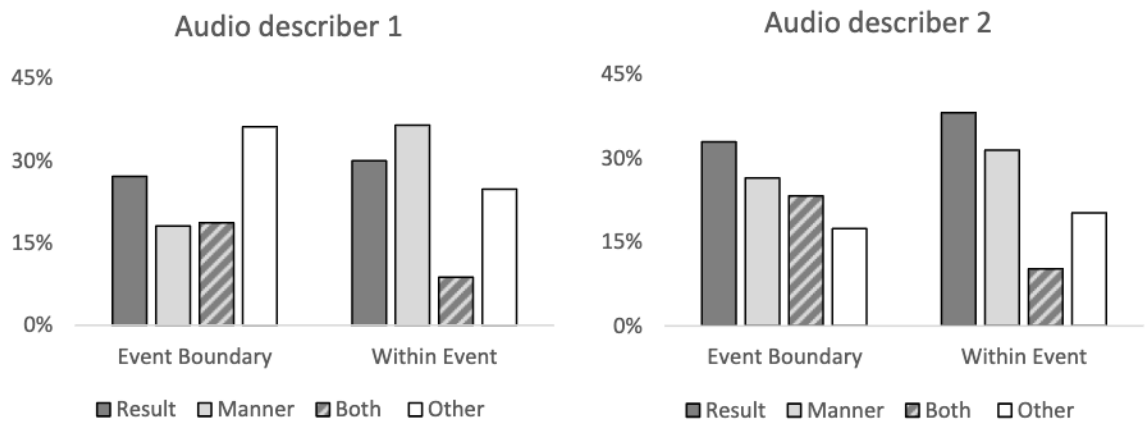
3.2. The Expression of Manner and Result

3.2.1. Expressions at Event Boundaries and Within Events

The proportion of Result and Manner expressions in relation to intonation units occurring at event boundaries and within ongoing events are illustrated in Figure 3.

Figure 3

Expressions of Result and Manner for intonation units at event boundaries and within events



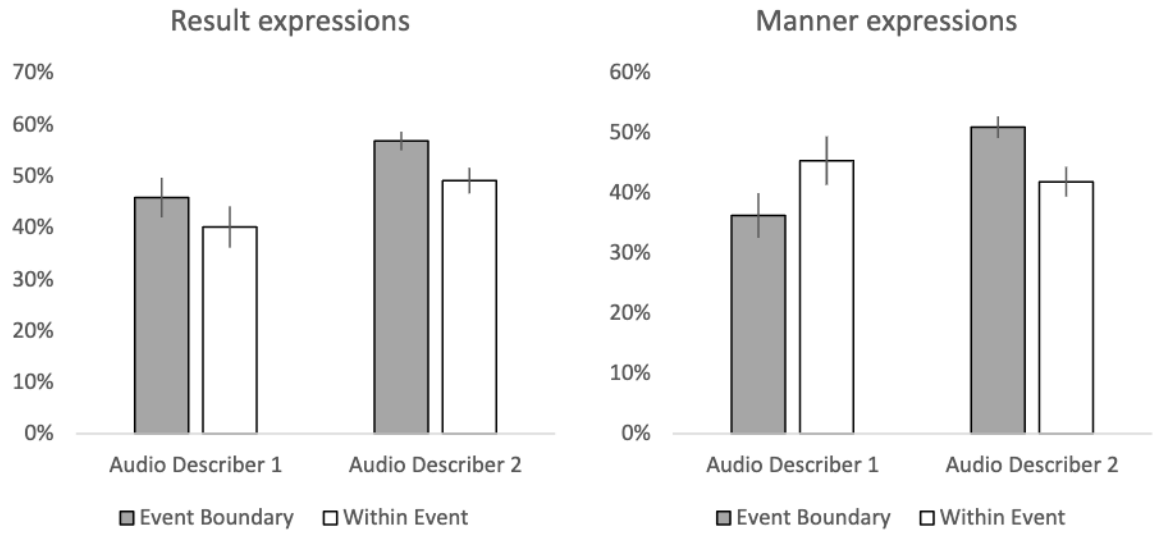
Note. Proportion of intonation units that included expressions of Result and Manner (or Both) for intonation units at Event Boundaries and Within Events. Other corresponds to expressions without any result or manner information.

Source: Authors' own study.

To statistically analyze to what degree Result and Manner were expressed at event boundaries, as compared to within events, we ran two Generalized Linear Models with intonation unit type (Event Boundary, Within Event) and Audio Describer as fixed effects (AD1, AD2). Results revealed that Result expressions were more common at event boundaries (p = .035) and that Audio Describer 2 used a higher proportion of Result expressions (p = .002). See Figure 4. For Manner expressions, no effects of intonation unit type or Audio describer emerged. But a significant interaction (p = .004) revealed that Manner expressions were more common within events for Audio Describer 1 (p = .03). See Figure 4.

Figure 4

Frequency of Result and Manner expressions at event boundaries and within events



Note. Frequency of Result and Manner expressions at Event Boundaries and Within Events over the two Audio Describers. Error bars denote SEM.

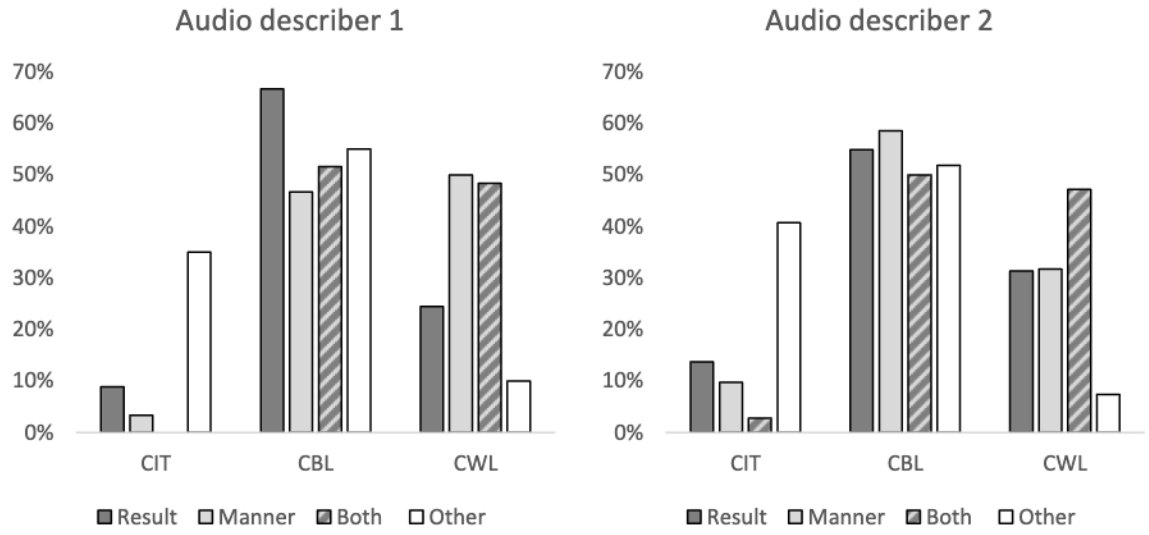
Source: Authors' own study.

3.2.2. Expressions at Event Boundary Categories

In the next step, we scrutinized how manner and result were expressed at event boundaries over the three categories. The proportion of Result and Manner expressions in relation to the three event boundary categories is illustrated in Figure 5.

Figure 5

Proportion intonation units with expressions of Result and Manner across three categories of event boundaries



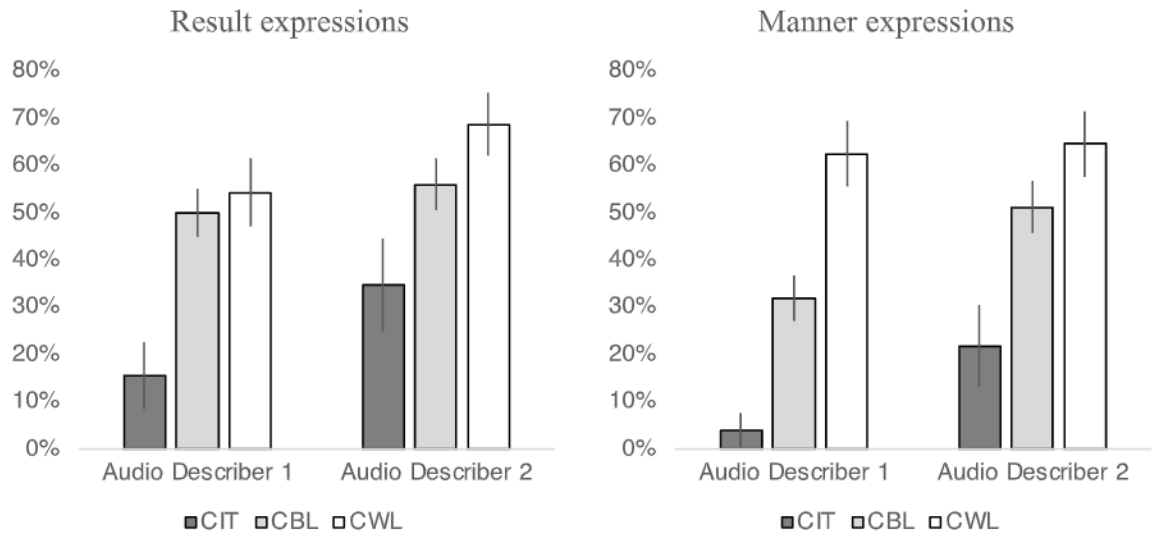
Note. Proportion of intonation units that included expressions of Result and Manner (or Both) across the three event boundary categories: change in time (CIT), change between locations (CBL), change within location (CWL). Other corresponds to expressions without any result or manner information.

Source: Authors' own study.

To statistically analyze to what degree Result and Manner were used across the three event boundary categories, we ran Generalized Linear Models with event boundary category (CIT, CBL, CWL) and Audio Describer as fixed effects (AD1, AD2). Results revealed that Result expressions were more common at CBL boundaries than CIT boundaries ($p < .001$) and at CWL boundaries than CIT boundaries ($p < .001$). No significant difference emerged between CBL and CWL boundaries. See Figure 6. For Manner, it was revealed that Manner expressions were more common at CWL boundaries than at both CBL and CIT boundaries ($p < .001$), and more common at CBL than at CIT boundaries ($p = .001$). See Figure 6.

Figure 6

Frequency of Result and Manner expressions between the two Audio Describers across three categories of event boundaries



Note. Frequency of Result and Manner expressions between the two Audio Describers across the three event boundary categories: change in time (CIT), change between locations (CBL), change within location (CWL). Error bars denote SEM.

Source: Authors' own study.

3.3. Expressing Event Boundaries in Audio Description

Different types of event boundaries were typically expressed with linguistically different resources. In this section, we exemplify and analyse some characteristic ways of expressing different types of event boundaries.

3.3.1. Change Between Locations (CBL)

CBL involved a high degree of Result expressions, but also a significant amount of Manner expressions. When Manner occurred in a Location change, it was typically expressed in the main verb together with Result in a particle or preposition. This type of construction is well attested in Swedish descriptions of spatial events (e.g., Blomberg, 2014), and is shown in (5) below.

5. En poliskonstapel springer in.
"A police officer runs in."

Another, more common strategy was to omit Manner and only express Result. A very common strategy for the audio describers was to use the verb *komma* ("come"), often together with particles and prepositions also expressing Result, see (6).

6. Här kommer hon till Stenkomst.
"Here she comes to Stenkomst."

When neither Result nor Manner were expressed, the audio describers typically used elliptical verb-less constructions with a new location specific. This would indicate that a change of location had occurred, such as in (7).

7. Ute på Alvaret.
"Out on Alvaret".

3.3.2. Change Within Location (CWL)

CWL was the boundary type most associated with Manner-information. However, these constructions were typically accompanied by an expression of Result as well, see (8).

8. Nils springer dit.
"Nils runs over there."

Sometimes Manner was expressed without Result. This was the case when the change was neither temporally nor spatially prolonged, see (9). In this description, there is motion, but no change of location is specified.

9. Julia kör i full fart längs vägen.
"Julia races along the road."
(lit. Julia drives at full speed along the road)

3.3.3. Change in Time (CIT)

Changes in time were expressed somewhat differently than locational changes. This was semantically reflected in the type of descriptions provided by the audio describers. A recurrent pattern was to describe temporal change coinciding with location change through copula constructions, see (10).

10. Nu är vi i Havanna.
"Now we are in Havanna."

As can be seen from this example, the audio describer uses the first-person plural pronoun, indicating that the situation is construed from the perspective of the viewers. The use of such viewer-centric construals occurred predominantly for temporal changes. Another common strategy was to just use a noun denoting a particular moment in time, e.g. (11) and (12).

11. Nästa morgon.
"The next morning."

12. Nästa dag.
"The next day."

Sometimes, the audio describers did not need to explicitly mention the temporal change, since the change in time could be inferred from the context. Examples of this involve the introduction of a character known by the audience to be present only in one of the three different time periods depicted in the film (13).

13. Martin vrider sig i våndor i sängen.
"Martin twists in agony in bed."

Changes in time thus typically exhibited a different pattern from changes in location.

4. Discussion

In the present case study, two independent audio descriptions (ADs) of the Swedish film *Skumtimmen*, that were produced live by two independent professional audio describers, were analysed and compared in relation to identified event boundaries in the film.

Results revealed that a noteworthy proportion of the audio describers' intonation units corresponded with and explicitly expressed spatiotemporal event boundaries (Audio Describer 1 = 18%, Audio Describer 2 = 28%). Moreover, most event boundaries were explicitly verbalised by both audio describers. Almost all event boundaries related to changes in time (CIT) and about 80–90% of the event boundaries related to changes between locations (CBL) were explicitly verbalised by both audio describers. Event boundaries concerning changes within location (CWL) were least likely to be explicitly verbalised (around 75% by both audio describers)[6]. All in all, these results demonstrate that event segmentation structure experienced from the film by the two audio describers is indeed also expressed in their corresponding ADs. This outcome is in line with the results by Gerwien and von Stutterheim (2018), who in a recent experimental study of language production, demonstrated that the perception of visually presented stimuli and spoken event descriptions of corresponding information to a large degree rely on the same event segmentation principles. However, as AD styles are known to vary both in respect to individual translators and films, the outcome of the present case study should primarily be considered as laying the groundwork for future studies examining larger samples of audio describers and films.

[6] An explanation of this particular result could be that changes within location (CWLs) are typically not marked by physical boundaries, and when there are no visual cues, such as doors and gates, the changes are less likely to get explicitly verbalised.

In contrast to previous language studies of event segmentation, where the *perception* of narratives has been targeted (e.g., reading or listening), the present study investigated event segmentation when language is *produced* in an AD context. A specific focus was on how the spoken discourse was construed in relation to manner and result information. When comparing intonation units at event boundaries with intonation units within events, it was revealed that result information was more frequently expressed at event boundaries. For manner information, the results were somewhat mixed, with only Audio Describer 2 including more manner information in the within event intonation units (and with a tendency for a reversed pattern for Audio Describer 1). Thus, as expected, discourse describing event boundaries was typically characterized by Result verbs expressing situation changes (cf. Warglien et al., 2012), whereas Manner verbs, expressing exertion of forces (cf. Warglien et al., 2012), were, at least for one of the audio describers, more common in discourse describing actions within an ongoing meaningful event.

Previous research has demonstrated that Result verbs tend to activate the dorsal pathway of the cortical visual streams, whereas Manner verbs tend to activate the ventral pathways of the cortical visual streams (Wu et al., 2008). The dorsal pathway is specialised for processing motion, spatial relationships, and action, whereas the ventral pathway is specialised for processing visual details (shape, colours, texture) and visual recognition (e.g. Mishkin et al., 1983). As the identified event boundaries in the present study were characterised by changes in the spatiotemporal context, it is conceivable that the dorsal pathway is frequently activated at event boundaries and then forming the basis for AD involving result information. However, as event boundaries to a large degree co-occur with the scene changes in the film, corresponding changes in visual content are likely to also activate the ventral pathway, which would then form the basis for AD involving manner information. This is consistent with a brain-mapping study of event segmentation during music perception (Sridharan et al., 2007) and with results from the present study showing a rather high degree of both result and manner information at event boundaries. The likelihood of activating both the ventral and dorsal streams at event boundaries could also explain the discrepancies in how manner information was used by the two audio describers in the present study, as there are documented differences in how individuals process and attend to visuospatial scene properties (Kozhevnikov et al., 2005).

When considering the three types of spatiotemporal event boundaries, results revealed that the audio describers used manner and result expressions to describe all three types of event boundaries, but to varying degrees. CBL and CWL were quite similar in respect to result information, but with significantly more manner information in CWL. This is expected, as both CBL and CWL are characterized by situation changes in the spatial "where-context" (cf. Wu et al., 2008), but where CWL are less abrupt and more likely to involve information about exertion of forces (e.g., running, crawling, rolling).

More surprising was the infrequent usage of Result- and Manner-expressions for CIT. This suggests that changes that occur in a spatial context are linguistically treated differently than changes in a

temporal context. It has often been argued that the semantic resources for talking about time are inherited from space (e.g., Lakoff, 1987; Langacker, 1987), to the extent that time is construed as a spatial concept (e.g., Casasanto & Boroditsky, 2008). Based on the way the audio describers expressed temporal changes, the validity of this assumption can be questioned – at least within the context of AD. Even though it is possible to use typically spatial language to talk about time, the ADs contained many indications of different strategies for construing time. One such indication was the relatively lower degree of Manner- and Result-expressions for temporal changes. The audio describers instead used expressions that did not take recourse to spatial constructions, such as temporal adverbials (e.g., *nästa dag* "next" day) or copula constructions (e.g., *det är morgon* "it is morning"), which is also consistent with what previous corpus-based studies of AD have reported (Jiménez Hurtado & Soler Gallego, 2013; Reviers, 2015; Salway, 2007). Thus, given the specific circumstances surrounding AD communication, where verbs are frequently not needed to indicate a change in the temporal framework, this result might not be that unexpected after all. However, how general this pattern is in AD should be studied further, taking into account different languages, as well as larger datasets of AD and audio describers over a wider range of films and tv shows.

Research on event segmentation and its verbalisation could be of great importance for the practice of audio description and training of audio describers, and has implications for the visually impaired end users (cf. Holsanova, 2016, 2022; Holsanova et al., 2020). Results of the present study could, for instance, make audio describers aware of the importance of verbalising event changes in relation to the spatiotemporal context and offer examples and solutions as to how this can be done. Event segmentation plays also an important role for automated, computer-generated video description (Braun et al., 2020; Starr et al., 2020). It is necessary to teach the algorithms to identify actions in dynamic scenes, to "see" connections between frames and actions, and to recognise event boundaries. Event boundary information will assist in developing algorithms that track referents across narrative evocations of time and space, which may offer support in establishing nominal and pronominal continuity and coherence (resolving some of the major issues in the performance of such algorithms). However, studies of AD also provide unique opportunities to directly investigate the interface between the perception of ongoing visual events and language production related to the corresponding information. As this interplay is virtually absent in current studies of event segmentation, we believe that more systematic research on audio descriptions could provide a prominent avenue for future studies on this understudied, but critical aspect.

The focus of the present case study was on the production perspective of AD: how the audio describer presents events in ongoing visual scenes verbally. However, there is also a need for corresponding research from a recipient perspective. How do users of AD understand and segment the described events? What event models do the end-users create, based on the verbal event description produced by the audio describers? How should critical event content be described so that the end users get involved and empathise with the story? What preferences do they have? It is important to conduct

interviews, explorative and experimental studies and to include and engage end users in the assessment of audio description (Holsanova, 2022).

References

Alfredson, D. (Director). (2013). *Skumtimmen* [Echoes from the dead]. Yellow Bird Films AB, Fundament Film AB.

Amoruso, L., Gelormini, C., Aboitiz, F., Alvarez Gonzalez, M., Manes, F., Cardona, J. F., & Ibanez, A. (2013). N400 ERPs for actions: Building meaning in context. *Frontiers in Human Neuroscience*, *7*(57).

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699.

Benecke, B. (2014). Character Fixation and Character Description. The Naming and Describing of Characters in *Inglourious Basterds*. In A. Maszerowska, A. Matamala & P. Orero (Eds.), *Audio description: New perspectives illustrated* (pp. 141-158). John Benjamins.

Blomberg, J. (2014). Motion in language and experience – actual and non-actual motion in Swedish, French and Thai [Doctoral dissertation]. Lund University.

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

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579–593.

Chafe, W. (1987). Cognitive constraints on information flow. In R. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21–51). John Benjamins.

Chafe, W. (1996). How consciousness shapes language. *Pragmatics and Cognition, 4,* 35–54. doi:10.1075/pc.4.1.04cha

Cutting, J., & Iricinschi, C. (2015). Re-presentations of space in Hollywood movies: An event-indexing analysis. *Cognitive science, 39*, 434–456.

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological science*, *22*(2), 243–252.

Firbas, J. (1992). *Functional sentence perspective in written and spoken communication*. Cambridge University Press.

Fresno, N. (2012). Experimenting with characters: An empirical approach to the audio description of fictional characters. In M. Carroll, P. Orero, & A. Remael, (Eds.), *Audiovisual translation and media accessibility at the crossroads.* John Benjamins.

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module]. https://gamlj.github.io/

Gärdenfors, P. (2000). *Conceptual spaces*. MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press.

Gerwien, J., & von Stutterheim, C. (2018). Event segmentation: Cross-linguistic differences in verbal and non-verbal tasks. *Cognition*, *180*, 225–237.

Givón, T. (1990). *Syntax: A functional-typological introduction*. Vol. 2 John Benjamins.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. Edward Arnold.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behaviour, 13*, 512–521.

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

Holsanova, J. (2008). *Discourse, vision, and cognition.* John Benjamins.

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

Holsanova, J. (2022). The cognitive perspective on audio description: Production and reception processes. In C. Taylor & E. Perego (Eds.), *The Routledge handbook of audio description* (pp. 57–77). Taylor & Francis.

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

Holsanova, J., Wadensjö, C., & Andrén, M. (2016). *Syntolkning – forskning och praktik* [*Audio description – research and practices*]. Lund University Cognitive studies 166/ MTM:s rapportserie nr 4.

Huff, M., Meitz, T. G., & Papenmeier, F. (2014). Changes in situation models modulate processes of event perception in audiovisual narratives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1377.

Jiménez Hurtado, C., & Soler Gallego, S. (2013). Multimodality, translation and accessibility: A corpus-based study of audio description. *Perspectives*, *21*(4), 577–594.

Johansson, R. (2016). Mentala bilder hos seende och blinda. In J. Holsanova, C. Wadensjö, & M. Andrén (Eds.). *Syntolkning – forskning och praktik* [*Audio description – research and practices*]. Lund University Cognitive studies 166/ MTM:s rapportserie nr 4, 29–38.

Johansson, R., Holsanova, J., & Holmqvist, K. (2013). Using eye movements and spoken discourse as windows to inner space. In C. Paradis, J. Hudson, & U. Magnusson (Eds.). *The construal of spatial meaning: Windows into conceptual space*, 9–28. Oxford University Press.

Klein, W. (2009). How time is encoded. In W. Klein & P. Li (Eds.), *The expression of time* (pp. 1–43). Walter de Gruyter.

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

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79.

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition*, *39*(1), 75–91.

Kvist Darnell, U. (2008): Pseudosamordningar i svenska: särskilt sådana med verben sitta, ligga och stå. (Pseudo-coordinations in Swedish, in particular with the verbs sit, lie and stand). Doctoral dissertation. Stockholm: Department of Linguistics. Stockholm university.

Lakoff, G. (1987). The death of dead metaphor. *Metaphor and symbol*, *2*(2), 143–147.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Descriptive application.* (Vol. 2). Stanford University Press.

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, *35*(6), pp. 1489–1517.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Mazur, I. (2015). Characters and actions. In Remael, A., Reviers, N., & Vercauteren, G. (Eds.), *Pictures painted in words: ADLAB Audio Description Guidelines*, EUT Edizioni Università DiTrieste, Trieste, Section 2.1.1, 19–23.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28–38.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition*, *41*(1–3), 47–81.

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

Rai, S., Greening, J., & Petré, L. (2010). A comparative study of audio description guidelines prevalent in different countries. *Londra: Royal National Institute of Blind People*.

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

Remael, A., & Vercauteren, G. (2015). Spatio-temporal settings. In: *Pictures painted in words: ADLAB Audio Description Guidelines*, EUT Edizioni Università DiTrieste, Section 2.1.2, 24–27.

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.

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.

Santin, M., Van Hout, A., & Flecken, M. (2021). Event endings in memory and language. *Language, Cognition and Neuroscience*, *36*(5), 625–648.

Skordos, D., Bunger, A., Richards, C., Selimis, S., Trueswell, J., & Papafragou, A. (2020). Motion verbs and memory for motion events. *Cognitive Neuropsychology*, *37*(5–6), 254–270.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(4), 335–345.

Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language*, *53*(1), 125–140.

Sridharan, D., Levitin, D. J., Chafe, C. H., Berger, J., & Menon, V. (2007). Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. *Neuron*, *55*(3), 521–532.

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.

Stawarczyk, D., Wahlheim, C. N., Etzel, J. A., Snyder, A. Z., & Zacks, J. M. (2020). Aging and the encoding of changes in events: The role of neural activity pattern reinstatement. *Proceedings of the National Academy of Sciences*, *117*(47), 29346–29353.

Talmy, L. (2000). *Toward a cognitive semantics*. MIT Press.

The Jamovi Project (2019). *Jamovi*. (Version 1.6.23) [Computer Software]. https://www.jamovi.org

Vandaele, J. (2012). What meets the eye. Cognitive narratology for audio description. *Perspectives*, *20*(1), 87–102.

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

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.

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.

Warglien, M., Gärdenfors, P., & Westera, M. (2012). Event structure, conceptual spaces and the semantics of verbs. *Theoretical Linguistics*, *38*(3–4), 159–193.

Wu, D. H., Morganti, A., & Chatterjee, A. (2008). Neural substrates of processing path and manner information of a moving event. *Neuropsychologia*, *46*(2), 704–713.

Zacks, J. M. (2013). Constructing event representations during film comprehension. In Shimamuram (Ed.), *Psychcinematics: Exploring cognition at the movies* (pp. 227–243), Oxford University Press.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651–655.

Zacks, J. M., & Magliano, J. P. (2011). Film, narrative, and cognitive neuroscience. In D. P. Melcher & F. Bacci (Ed.), *Art & the senses*. Oxford University Press.

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

Zacks, J. M., Speer N. K., Swallow K. M., Maley C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience*, *4*(168), 1–15.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin*, *127*(1), 3.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292–297.

Zwaan, R. A., Magliano, J. P., &Graesser, A. C. (1995). Dimensions of situation-model construction 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.