Reviers, N. (2018). Tracking multimodal cohesion in Audio Description: Examples from a Dutch audio description corpus. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *14*, 22–35.

Tracking multimodal cohesion in Audio Description: Examples from a Dutch audio-description corpus

Nina Reviers

University of Antwerp, Department of Applied Linguistics/Translators and Interpreters, Trics research group, Belgium
Nina.reviers@uantwerpen.be

One of the main questions addressed by multimodality research—the main conceptual framework for analysing audiovisual texts—is how the different modes of audiovisual texts combined—visual, verbal, aural—create supplementary meaning in texts, over and above the meanings conveyed by the individual constituents. Ensuring that this multimodal interaction or multimodal cohesion remains intact is a key challenge in the practice of audiovisual translation (AVT), and particularly in Audio Description (AD) for the blind and visually impaired. The present article therefore studies the functioning of multimodal cohesion in audio-described texts by analysing the types of interaction between descriptive units and sound effects in a selection of Dutch audio-described films and series. The article begins with a detailed description of the methodology which is based on multimodal transcription and concludes with an overview of the types of multimodal cohesive relations identified.

1. Introduction

Multimodality—the principle that texts convey meaning through the combination of verbal, visual and aural–semiotic modes—intrinsically sits at the core of any study of Audiovisual Translation (AVT). Multimodality is fast becoming the main conceptual framework for the study of audiovisual texts and their various translations (Pérez-González, 2014). While it may be obvious to approach the study of Audio Description (AD) from a multimodal perspective, developing an adequate theoretical and analytical framework to do so is another matter entirely. At present, the multimodal tradition in text analysis is still an emerging paradigm drawing on an eclectic collection of studies from a range of different disciplines.

Multimodality research focuses on describing the meaning-making process of verbal as well as non-verbal modes such as images and sounds by inventorizing their possible resources and their semiotic potential (Van Leeuwen, 2005, p. 4). One of these resources is the interaction between and across modes, which creates supplementary meanings in addition to those conveyed by each of the modes individually. Baldry and Thibault (2006, pp. 7–19) have coined the term "resource integration principle" to describe this phenomenon. Royce (2007, p. 63) refers to this process as "intersemiotic complementarity" and Van Leeuwen (2005, p. 179) describes a similar phenomenon in the field of Social Semiotics with the term "multimodal cohesion", which he defines as "the integration and co-occurrence of different kinds of semiotic resources". Terminological differences aside, the above scholars agree that the combination of different modes is the most powerful feature of the meaning-making process of multimodal texts—including audiovisual texts—and it is only through their interaction that the meaning potential of individual modes is materialized.[1]

In the field of AVT and Media Accessibility (MA) in particular, this process is also a key factor. One of the main challenges of AVT practice is to ensure that cohesive links between modes in the source text (ST) remain intact in the target text (TT). This is particularly challenging in the case of AD. AD is a form of MA that offers a verbal description of the relevant elements of the visual component of a work of art or media product to those unable to perceive it themselves, so that all audiences can fully grasp its form and content. When audio-describing audiovisual texts, relevant links between the visual mode and sounds—be it music, sound effects or voice timbre—are to be recreated or compensated for in the audio-described version in order to maintain the text's overall cohesive strength and to ensure that the narrative structure remains intact (Braun, 2011; Zabalbeascoa, 2008). The way this "multimodal cohesion", following Van Leeuwen's terminology, functions in audiovisual texts and their translations is complex and research has not yet revealed the full potential and understanding of this meaning-making instrument.

To date, multimodal cohesion has been approached from a range of different angles using different terminologies, however, a common conceptual framework to study the phenomenon does not exist. What is more, current research tends to focus on the visual aspects at the expense of the role of the aural mode, particularly music and sound effects, in multimodal meaning-making. The semiotics of sound, however, is particularly relevant to AD research as the AD text is a predominantly aural text type in which dialogues, music and sound effects jointly create meaning. In view of this, it is striking to note how little attention has been devoted to the role of sound in AD. The Ofcom guidelines (ITC, 2008, p. 18), to give one example, acknowledge that description is necessary for "any sounds that are not readily identifiable". However, few studies elaborate further on how to identify "important" or "unidentifiable" sounds or how sounds complement the descriptive units of AD to convey the entire message of the text (see also Orero & Szarkowska, 2014).

In brief, addressing the topic of multimodal cohesion in AD raises a series of convoluted questions:

- How can multimodal meaning-making and, in particular, multimodal cohesion be studied in audiovisual texts and AD?
- What is the role of sound in the meaning-making process of AD?
- How is multimodal cohesion between sound and AD created at a textual level?

These questions were explored in a four-year PhD project conducted at the University of Antwerp in Belgium.[2] The project combined quantitative and qualitative (multimodal) corpus methods to analyse the typical characteristics of the language use in AD. The final stage of the project adopted a multimodal approach to the study of AD and the results of this stage are discussed in the present article (previous results are presented in Reviers, 2016; Reviers, Remael, & Daelemans, 2015).

Section 2 of this article presents the theoretical and methodological framework developed for analysing multimodal cohesion in AD. I discuss the method of multimodal transcription, which forms the basis for the analysis, and the literature consulted to develop an adequate theoretical and terminological framework. Section 3 reports on the analysis of the functioning of multimodal cohesion in a selection of audio-described texts taken from the corpus of Dutch ADs that was developed as part of the PhD project underlying this article. The section offers a summary of the findings of the PhD project, focusing on the types of multimodal cohesion identified and the role of sound in them.

2. Theoretical and methodological framework

In the tradition of multimodality research, Baldry and Thibault (2006) have developed the method of multimodal transcription, which is becoming a much-used methodological tool in multimodality research into both original and translated texts (Hirvonen & Tiittula, 2005; Pérez-González, 2014; Taylor, 2003). The central idea of multimodal transcription is to segment the audiovisual text into its smallest constitutive units or modes by transcribing them in a column-based table in order to facilitate objective and empirical analysis. Each column represents a different textual and/or analytical layer in accordance with the analytical goals being pursued by the research project. In other words, Baldry and Thibault's system is composed of a few basic analytical principles that can be adapted according to the needs of the research project (Pérez-González, 2014). Multimodal transcription can be broadly summarized according to four basic principles (Baldry & Thibault, 2006): Sections 2.1 to 2.4 below offer an overview of them. They are supplemented by a discussion of the individual transcription needs related to the study of sound in AD—as multimodal transcription is an innovative instrument in this respect—and the theoretical framework used for adapting the model to these research needs. Section 2.5 presents the actual transcription table and its annotation set.

2.1 Principle #1: Identify salient features of the semiotic modes

The transcription of semiotic modes is not based on the notation of all the physical criteria of a mode, but highlights those features that are perceptually and semiotically most salient. When it comes to analysing sound, it cannot be reduced to the study of acoustic properties such as loudness or pitch, but should also include the transcription of the *effects* and *types* of meaning of sound (Baldry & Thibault, 2006, pp. 180–210). But what relevant types of meaning and effect can the soundtrack of an audiovisual text convey?

According to Chion (1999), a key figure in the study of the role of sound in film, a film's aural and visual modes are extremely closely connected, to the extent that it is difficult to speak of one without mentioning the other: "*On ne 'voit' pas la même chose quand on entend; on 'n'entend' pas la même chose quand on voit*" [we don't see the same thing when we also hear; we don't hear the same thing when we also see] (p. 3). This multimodal interaction forms the core of the filmic medium and is what Chion calls the "audio-visual contract". Sound has long been thought to play a secondary role in this partnership because audiences are not always aware of the added value of sound. They almost always interpret sounds through the visuals by automatically associating it to what is happening on screen (Bordwell & Thompson, 2013; Chion, 1999; Fryer, 2010). This raises a few fundamental questions regarding the study of AD: To what extent can sounds be clearly identified by the audience without the support of a visual?; and Do sounds carry the same semiotic value in the audio-described version? Given these observations, the relevance of the effective integration of sounds and descriptions becomes even more apparent in AD.

The role of the soundtrack is taken into account more often in multimodality and film studies, and a body of literature that analyses sound as a semiotic resource in its own right is starting to take shape (Bordwell & Thompson, 2013; Chion, 1999; Van Leeuwen, 1998, 2005). In these works, the following key properties of sound surface: the types of sound, the acoustic properties of sound and the effects sounds conjure up.

First, there are different types of sound and terms to identify them. The relevance for the analysis of AD is that sounds are usually classified according to the source of the sound, given the audiovisual contract mentioned earlier. This accounts for why AD literature and guidelines advise describers to identify primarily the source of sounds that are not readily identifiable (ITC, 2008; Remael, Reviers, & Vercauteren, 2014). Sounds are categorized either according to the position of the source in the story world (diegetic)

or outside the story world (non-diegetic), or according to their position on or off screen. Sounds can also be categorized according to the nature of the source: sounds synchronized with the actions of an on-screen character (foley sounds), sounds (re-)creating the sound of a location or a setting (ambience sounds), sounds that add a connotative layer to a scene (impact sounds), and even silence (Chion, 1999; Turner, as cited in Remael, 2012). Finally, there are also sounds that are heard without their sources being known or being visible on screen. Chion (1999, pp. 63–65) calls these "*sons acousmatiques*" or acousmatic sounds.

Secondly, Chion (1999), Van Leeuwen (1999), and Bordwell and Thompson (2013) discuss various properties of sound that filmmakers can actively draw on and manipulate in order to create different effects, such as loudness (volume), pitch (the perceived highness or lowness of a sound) and timbre (the colour or tone quality of a sound). These acoustic qualities are created or manipulated in post-production and many sound effects bear little resemblance to their supposed source in real life (e.g., coconuts being used to imitate the sound of hooves). As Remael (2012, p. 261) points out,

> [F]ilm sound is anything but reproductive and has surpassed the area of indexicality, i.e. there is not a straightforward relationship between a given film sound and a sound that exists in a pre-production environment.

Film-makers do not simply select an appropriate type of sound. They actively manipulate its properties to heighten certain qualities and highlight its narrative function. This raises the question to what extent film sounds and their sources can easily be identified in the described version? On the other hand, film sounds are selected and manipulated to enhance their narrative function (Bordwell & Thompson, 2013). Contrary to sounds in real life that blend in randomly with other environmental sounds, the narrative focus given to filmic sounds might actually enhance the audience's ability to interpret sounds correctly.

To conclude this section, I want to stress that the types of sound and their properties give film-makers the liberty to create effects to complement a film's visual mode and overall narrative. The effects that can be created in this way are, however, theoretically endless in number and, therefore, film-makers constantly experiment with and generate new effects through sound. As a result, it becomes impossible to compile an exhaustive list of the possible effects or functions of sound in film. A few common effects discussed in the literature are the creation of realism, the creation of a real-time effect, sound perspective (discussed in the next section) and the creation of impressionistic or symbolic effects by manipulating the "modality" of sounds, that is, whether the sound quality is abstract or naturalistic or has a high sensory impact (Van Leeuwen, 1998).

In order to take these effects into account systematically in my analysis, I apply a classification of AD presented by Fryer (2010): it constitutes a useful bridge for transferring the insights from Films Studies and Social Semiotics presented in the previous paragraphs to the study of sound in AD. Fryer (2010) approaches the issue from the perspective of audio drama and defines sound effects in relation to the dialogues, or according to the effect they create by themselves. It seems that most of the effects of sound in film described in the literature (by Bordwell & Thompson, 2013; Chion, 1994, 1999; Van Leeuwen, 1999, 2005) can be grouped into the six categories Fryer (2010) uses:

- the realistic confirmatory effect (when a sound corresponds to a piece of information also mentioned elsewhere in the film text);
- the realistic evocative effect (when only the sound is the source for identifying a piece of information);
- the symbolic effect (when sounds are used to symbolize a piece of information, for example a recurring sound effect every time the murderer commits a crime);

- the conventionalized effect (for common sounds from everyday life, e.g., a mobile phone ringing);
- the impressionistic effect (when a sound has an emotional impact), and
- music as an effect (when music is used to convey the effect).

2.2 Principle #2: Segment the text into phases

The second principle of multimodal transcription is that the vertical breakdown of the text into rows in the multimodal transcription table is based on the segmentation of the text into phases. A given phase "is characterised by a high level of metafunctional consistency or homogeneity" (Baldry & Thibault, 2013, p. 47).

According to Baldry and Thibault (2013), the Prague school concept of foregrounding is crucial when showing which selections from which semiotic resource system are relevant for the instantiation of a given phase. The concept of foregrounding has been applied regularly to the study of the visual mode, but Van Leeuwen (2005) has applied the same principle to the analysis of sounds.

Sounds have the ability to create a sense of distance or proximity between a given object and its viewer or listener. This distance can be interpreted as physical distance, spatial orientation or narrative relevance. In the case of narrative relevance, a sound can be in "figure" position, when it is treated as the most important sound through properties of volume, for instance. Or a sound can be in "ground" position, when it is less perceptually salient and forms the background against which the action takes place (Van Leeuwen, 1998).

A second parameter to be considered is rhythm: "a transition from one textual phase to the next can be expected to relate to the overall rhythmic patterning of the text in significant ways" (Baldry & Thibault, 2006, p. 47). The transcription of rhythm should attempt to reveal the synchronicity between modes and their organizational patterns by, for instance, indicating accented rhythmic units or "pulses", signalling rhythmic groups, or indicating the degree of loudness, duration, tempo or pauses.

What is of particular interest in the present study is that rhythm is also highlighted by Van Leeuwen (2005) as an important parameter of multimodal cohesion. Rhythm is seen to provide coherence and meaningful structure to events that develop over time, and it is one of the single most important sources of cohesion in audiovisual texts. Van Leeuwen (2005) developed a series of parameters for analysing time and rhythm that are applied in the analysis in this article and will be exemplified in section 3.

2.3 Principle #3: Analyse the interplay between modes in and across phases

In order to analyse the interplay between modes across phases in longer stretches of film, Baldry and Thibault (2006, pp. 232–234) introduce the concept of "participant chains": the cross-modal repetition of participants (such as characters, objects or settings) across shots and phases. This concept has been further developed by Tseng (2013). Tseng's work focuses on developing an analytical framework for narrative films and is rooted in the same Social Semiotics tradition as Royce (2007) and Van Leeuwen (1998, 2005) mentioned earlier. Tseng (2013) demonstrates that film viewers draw on four types of element for constructing a coherent film narrative: characters, objects, settings and action. She argues that "It is the cohesion of the characters' and objects' identity tracking [through cohesive chains] which plays the significant role in guiding the path of narrative interpretation" (Tseng, 2013, p. 82). Tseng's (2013) model for analysing filmic cohesion follows five steps or "systems":

- To begin with, there is the system for presenting characters, objects and settings, which can be done mono- or cross-modally, and with either immediate or gradual salience.
- This is followed by three methods for tracking the way in which participants can be re-identified throughout the remaining text. These are what Tseng (2013) calls the "presuming system" (whether an object or character reappears explicitly or implicitly), the "comparative system" (whether an object or character reappears similarly to or differently from previous appearances) and the system for determining the direction of the identity retrieval (including whether it appeared earlier or will appear later, whether the identity retrieval is explicitly or implicitly cued and whether it is expressed mono- or cross-modally).
- The final step deals with an analysis of the action patterns across the film, which function as a cohesive tool between the characters that perform them, the objects used in them and the settings in which they take place.

2.4 Principle #4: Integrate a metafunctional analysis

One of the aims of multimodal transcription is "to describe short sequences of dynamic video texts in terms of the relationship between phases and metafunctions" (Baldry & Thibault, 2006, p. 167). The term "metafunction" comes from Social Semiotics and Systemic Functional Linguistics, which form the roots of multimodality research, and are used to analyse the different levels of meaning on which language (both verbal and non-verbal) can operate (for more, see Halliday & Matthiessen, 2014). While the metafunctional analysis of sound in AD is beyond the scope of the present article, one central concept of this approach is worth mentioning.

Royce (2007) is one of the authors who has taken this approach regarding what he calls "intersemiotic complementarity" (see introduction). He explains that multimodal interaction can be conceptualized in terms of what is called "lexical cohesion" in traditional linguistics: repetition, for the repetition of a piece of information by two different modes; synonymy, for a similar type of meaning expressed by different modes; antonymy, for an opposite type of meaning; hyponymy, for the classification of a general class of something and its subclasses; and meronymy, for reference to the whole of something and its constituent parts.

This idea is very closely related to the concept of "information-linking" introduced previously by Van Leeuwen (2005) as a key parameter of multimodal cohesion. This term refers to the cognitive links audiences can potentially construct between items in terms of causal or temporal relationships. Royce's (2007) concept of lexical cohesion can be seen as one type of information-linking, but previous research has indicated (Remael & Reviers, 2018) that the concept of "information-linking" also covers more implicit co-occurrences of items. These are often created by what Chion (1999) calls "synchronization points", when two elements in the text are synchronous and therefore that they are explicitly linked to each other in the audience's mind. The process of information-linking, implicitly through the co-occurrence of items of information or more explicitly by the creation of synchronization points, also plays a key role in the cohesion of audiovisual texts.

2.5 Multimodal transcription model for the analysis of AD texts

Based on the framework presented above, a selection of fragments from the Dutch AD corpus has been transcribed. Table 1 includes the transcription of the opening scene of the Flemish film *Loft* (Van Looy, 2008), which will serve as an example to illustrate the transcription method.

1.Phase	2. Descriptive unit	3. Figure	4. Ground	5. Time	6. Identity-tracking	7. Effect
	Macro-phase 1					
1	(4907) Een modern flatgebouw doemt wazig op in een blauwige nevel. Het is nacht. [A modern apartment building appears fuzzily in a blueish fog. It is night.]		Non-diegetic sound (mechanical, breathy drone) +Ambience (*rain*)	Time: unmeasured-fluctuating, slow SP: pulse in non-diegetic sound + AD ("doemt op" [appears])	Objects: generic - monomodal - immediate salience Setting: generic - cross-modal - gradual salience Action: conceptual process - cross-modal	Symbolic effect of SP +Realistic, evocative effect ambience (low naturalistic modality) +Impressionistic effect of non-diegetic sound and ambience (high sensory modality) = suspense
			↓ Non-diegetic sound and ambience continues	↓ Time continues	Indirect continuation of setting (rain)	↓ Impressionistic and evocative effect continue
	(4908) De terrassen van het gebouw zijn verlicht. Woestijnvis presenteert. [The terraces of the building are lit. Woestijnvis presents.]		↓	↓	Direct, monomodal, different reappearance object Direct, continuation of same setting (rain) + Action: conceptual process - monomodal	↓

			↓		Direct, continuation of same setting (rain)	↓
2	(4909) Architectonische lijnen en spaarzaam licht glijden over elkaar heen. Met de steun van het Vlaams Audiovisueel Fonds en Een. [Architecural lines and dim lights glide over each other. With the support of the Flemish Audiovisual Fund.]		↓ + Non-diegetic music Non-diegetic (“swoosh”)	SP: “swoosh” when non-diegetic sound and music overlap while AD reads “glijden over elkaar” [glide] Time: music unmeasured, fluctuating	Indirect, monomodal, different reappearance of Object (hyponymy) Direct, continuation of same setting (rain) + Action: conceptual process - cross-modal	+Symbolic effect of SP (high abstract modality) +Impressionistic effect of music and non-diegetic sound (high sensory modality) ↓ Evocative effect of rain continues

Table 1: Multimodal transcription of the opening scene of *Loft* (Van Looy, 2008)

Column 1 divides the scene into phases and sub-phases (following Principle #2 described above) and column 2 includes the transcription of the AD. Columns 3 and 4 include the annotation of the types of sound and their salience (whether a sound is situated in figure or ground position and whether it moves from one position to another. Column 3 in the present transcription is empty, as no sound occurred in figure position). The salience of a sound is determined based on its acoustic properties such as loudness, pitch and timbre, following Principle #1 discussed above.

In Table 1, for example, the first phase consists of two descriptive units in figure position (column 2). In ground position (column 3) a non-diegetic sound is featured, namely, a low, tense, mechanical, breathy, droning or humming sound that is not created by a source in the story world. This non-diegetic sound effect is combined with an ambience sound, namely, the sound of rain. The rain is an acousmatic sound, because the source of the sound is not mentioned in the AD (it is mentioned much later on). These two sound effects continue across both descriptive units (indicated by downward arrows in the table). Phase 1 is linked to phase 2 by the continuation of the non-diegetic and ambience sound effects in ground position, which indicates that both phases are part of the same macro-phase. A transition to a new phase is signalled by the addition of another non-diegetic sound effect, namely, a musical score in ground position. We can also hear a non-diegetic “swooshing” sound, which occurs in ground position because of its relatively low volume. After the first descriptive unit ends and the second unit starts, an additional diegetic, ambience sound effect is mixed in with the other sound effects in

ground position, namely, the distant sound of a police siren. Just like the rain, it serves as an acousmatic sound, since the police siren is only explicitly heard and mentioned later in the scene, when it moves from ground to figure position as its volume grows louder and louder.

Column 5 includes the annotation of issues related to timing and rhythm (see Principle #2 above). This includes explicit "synchronization points", following Chion's (1999) term (see section 2.1 on Principle #1 above). In Table 1, for example, both descriptive units of phase 1, the non-diegetic sound and the rain, are presented simultaneously, that is, they are placed in the same row. The timing of the sound stream created by these aural modes can be labelled as "unmeasured" and "fluctuating" following Van Leeuwen's (2005) classification, that is, they have no clear beat or rhythm. The simultaneous timing of the descriptive unit and the non-diegetic sound effect creates an explicit synchronization point (indicated with the abbreviations SP in column 5): just as the AD voice reads the words "doemt wazig op" [appears fuzzily], a pulse occurs in the fluctuating non-diegetic sound, which accentuates the words that are described.

Cohesion is also cued by the way in which characters, objects, actions and settings are cross-modally identified within and across phases (see Principle #3 discussed above). Following Tseng (2013), I therefore annotate how these chains are materialized in column 6. The consecutive descriptive units in Table 1 identify an object (a building), a setting (nighttime, rainy) and an action (appearing). In Table 1 only the identification chain of the building is indicated in bold typeface, as an example.

The first descriptive unit presents a new object monomodally (only mentioned in the AD), namely, with the words "Een modern flatgebouw" [a modern apartment building]. The object is presented as generic, since it is new and unknown, as indicated by the indefinite article "een" [a]. It is presented with immediate salience, since it is put in theme position as the subject of the sentence and not presented gradually. In the second descriptive unit, the building makes a direct, monomodal reappearance as the word "gebouw" [building] is repeated in the first sentence of the descriptive unit (lexical relation of repetition, following Royce's (2007) logic, see Principle #4 above). In terms of Tseng's (2013) "comparative system", the object reappears differently in quality, since additional features of the building are revealed through the lexical relation of meronymy (the building has terraces).

According to Tseng's (2013) "presuming system", which is used to re-identify participants, there is the indirect, monomodal reappearance of the building in phase 2 by means of lexical relations in the AD: the "architectonische lijnen" [architectural lines] refer to the building by way of hyponymy.

Column 7, finally, annotates the effects of the sounds, which are determined based on the acoustic qualities of the sound (loudness, pitch and timbre) and the level of "modality" it creates (see Principle #1 in section 2.1).

The overall effects created this way are labelled using Fryer's (2010) categorization introduced in section 2.1. For example, the synchronization point in phase 1 between the descriptive unit and the non-diegetic sound effect (see column 5) suggests a semiotic relation that audiences can infer through the process of information-linking. In particular, the sound effect is a *symbolic* aural representation of the action of "opdoemen" [appearing], an action that has no realistic aural equivalent in real life and is represented through an artificial, non-diegetic sound, as if the action of "opdoemen" would create a sound. The ambience sound (rain) in phase 1, in turn, creates a realistic, evocative effect—realistic because it helps to set a realistic location, and evocative because the rain is not mentioned elsewhere, so the sound alone evokes the idea of rain. It is important, however, to mention that the sound of rain has a low naturalistic modality (it is not the typical sound of raindrops, but a gushing sound). This raises the question whether audiences will be able to recognize the sound and whether it is indeed evocative. Finally, the non-diegetic sound in phase 1 and the ambience sound of rain also create an impressionistic effect. The music is meant to create a sense of suspense and impending

danger. It is a clear example of what Fryer (2010) labelled "music as an effect" and demonstrates that sounds are often used to illicit an emotional or affective reaction.

3. Analysis

The multimodal transcription model for AD developed above has provided a solid basis for analysing three audio-described fragments selected from the Dutch AD corpus (from *Loft*, a film of 2008 directed by Erik Van Looy, *Wolven*, a series developed by the Flemish Public Broadcaster VRT in 2012, and *Ben X,* a 2007 film directed by Nic Balthazar). The aim of the analysis is to identify more clearly how sound and AD work together to facilitate the (re)creation of cohesion in AD and those lexico-grammatical features that can be used to support this cohesion. The present section summarizes the main observations drawn from the exhaustive analysis conducted in the PhD project underlying this article about the types of multimodal interaction observed.

The way interaction between AD and sound is achieved in the three scenes under study varies on a scale indicating different degrees of explicitness, some being very explicit, others remaining more implicit. Three types of sound seemed to have a direct relation to the events described: realistic, confirmatory sound effects (naturalistic sounds which refer to a source that is also referenced in the dialogue or the AD); acousmatic sounds (more particularly sounds of which the source is only mentioned later by the dialogue or descriptive unit), and sounds with an impressionistic effect. The multimodal analysis revealed four ways in which the descriptive units refer to or disambiguate these types of sound implicitly or explicitly. These four ways are described below and illustrated with a few examples.

First, the descriptive unit *mentions or reiterates* the source of the sound *directly or explicitly* through the use of nouns or verbs that refer to it (examples 1 and 2):

(1) The description "Flashing lights approach" disambiguates the sound of sirens in the opening scene of *Loft*.[3]
(2) The AD from *Wolven* reads: "The stairs in front of the Museum of Modern Art. In the dark, numerous torches in large stone balls give the surroundings a fairy-like look. It is night time and a bunch of fancy people in dark suits and evening dresses is talking outside. Attendants in green vests [waistcoats] help out with incoming vehicles." The nouns and verbs in this description disambiguate the foley sounds of cars passing by and the ambience sounds of voices, footsteps and the ruffling of evening dresses.

Secondly, the descriptive unit *evokes* the source of the sound *indirectly or implicitly* through a process of lexical relations. In particular, the relation of meronymy (part–whole relation) seems common in AD (examples 3 and 4):

(3) In *Wolven* the description "They go inside" disambiguates the sound of a door opening, a bell ringing and footsteps.
(4) In the film *Ben X,* the AD at one point mentions that the characters are in the kitchen, which disambiguates the sounds of glasses and silverware that can be related to the kitchen through the lexical process of meronymy.

Thirdly, the descriptive unit (in addition) refers to the *quality* of the sound rather than the source to disambiguate it further and it uses nouns, verbs and, occasionally, adjectives to do so (example 5):

(5) The AD in *Loft* of "The rain gushes over the hood of the car" helps audiences to disambiguate the sound of rain, since the rain is not simply falling down, but gushing down hard. Without the verb referring to the sound quality, the sound might not be as easily identifiable due to its low naturalistic modality.

Fourthly, the descriptive unit indirectly and implicitly supports the process of *information-linking*, based on which audiences can infer the source of the sound (examples 6 and 7):

(6) In *Wolven* at one point a metallic, clinking sound can be clearly heard in figure position. Neither the AD nor the dialogues refer to the source of the sound. Audiences have cognitively to link the sound to an as yet unidentified object in a small, dark bag described in the AD and a hologram mentioned in the dialogue in earlier scenes.

(7) In the opening scene of *Loft*, the building mentioned earlier in this article is identified as a rooftop terrace with a railing. These are relevant features when one takes into account the world knowledge participants bring to the text. It suggests that the person who fell from the roof in a previous description fell from a great height and is likely to be dead. It also signals potential malicious intent, since it is not usual to fall accidentally from a rooftop, especially when there is a railing around the terrace (a piece of information that the ominous non-diegetic music also suggests).

Finally, the analysis identified two parameters that support the degree of explicitness of the above relations between sound and AD—since whether or not a sound is categorized as explicit depends on different criteria—namely, timing and sound quality. Timing relates to the degree of synchrony between the descriptive unit and the sound effects, that is, when they occur simultaneously or near-simultaneously the relation is more explicitly cued by the text. When the time between the items is extended, the relation is implicit. Consider example 8:

(8) In *Loft* the ambience sounds of rain and the police siren can be heard, but their sources are not identified by the AD from the start. Only a few moments after their first appearance are the rain and the police siren mentioned. In this example, the sound quality of the rain also creates a more implicit cohesive strength, due to its low naturalistic modality.

Timing is also the basis of rhythmical patterns, and when the descriptive units are deftly timed in with the rhythm of the soundscape, their interaction can be more easily determined by the audience (example 9):

(9) In the selected scene from *Wolven* two police officers chase after two criminals. This action scene as accompanied by fast, up-tempo music around which the staccato and quick AD is deftly timed. At a crucial moment the rhythm changes, contributing semantically to what is being described: one of the criminals pulls a gun on the other and they freeze. Simultaneously, there is a pause in the music and the beat is briefly replaced by fluctuating, choir-like female voices, an effect that seems to stretch the sense of real time. When the beat returns and the tempo picks up again, the men "unfreeze" and start fleeing from the police officers.

Other sound variables that can heighten the explicitness of the relation between sound and AD are volume, pitch and timbre, which put a sound either in figure or in ground position. An example is the sound of the siren in the opening scene of *Loft*, which first figures in ground position and is barely audible. Later, the sound gradually grows louder as it gains narrative salience and moves into figure position. Simultaneously, the sound becomes more explicit and can be more easily recognized.

Finally, sounds contribute meaning in audio-described texts in their own right even without support from the AD, either implicitly or explicitly. Conventionalized sounds and sounds with a high naturalistic modality are explicit sounds that are not usually described because they are considered to be readily identifiable. Another type of explicit sound is impressionistic sound effects, including non-diegetic music and sound effects and subjective sound perspective. These are sound types that appear relatively frequently without explicit disambiguation by the other aural modes, including the descriptive units, in the scenes under analysis. Consider examples 10 and 11.

(10) In *Ben X*, sound is the sole source of information conveying the subjective perspective of the main character. Certain foley sounds are unrealistically loud, placing them in figure position and creating a high sensory effect. Examples are the sound of a clock ticking or the sound of Ben's knife scratching the plate as he carefully cuts his bread. The volume of the sounds is used to reflect Ben's subjective aural perspective: as a boy with autism, he perceives these sounds more loudly than others. Other sounds are placed in the background, such as the volume of his mother's dialogue, just when Ben shifts his focus from his mother to cutting the slice of bread on his plate.

(11) The non-diegetic music in the opening scene of *Loft* contributes to the sense of impending danger independently of the AD, which does not explicitly mention danger or suspense.

The multimodal analysis also identified the presence of sounds that are less readily identifiable (sounds with a low naturalistic modality, low-volume sounds and sounds which are not conventionalized), but that were nevertheless not disambiguated by the descriptive units (or the other aural modes). These sounds therefore run a higher risk of creating incoherence and should be addressed in audio-described texts as far as possible when time allows (example 12).

(12) At one point in the opening scene from *Loft*, a female character refers to a woman by using vague and ambiguous terms: "her", "victim", "that sense". The entities to which these words refer are only identified by a foley sound, namely the ruffling sound of papers. As a result, the foley sound is expected to have a realistic, evocative effect as it is the only semiotic resource helping audiences to identify the source of the words. The information-link the audience is supposed to make here is difficult and they need to rely to a large extent on their background knowledge of the text and the narrative context in order to infer what is happening: the ruffling sound of papers is coming from a stack of photos the female detective is holding on which the victim is depicted.

4. Concluding remarks

The analysis of a selection of audio-described scenes from a multimodal perspective has confirmed that sounds contribute greatly to creating a rich and vivid TT. It also confirms

that the skilled integration of sound and AD can increase the cohesion of the text at different levels, moving from more implicit to more explicit strategies. This supports the view that research should attach more importance to the study of sound, not only in film studies and multimodality research, but also in AVT and MA.

By combining the principles of multimodal transcription with the various concepts regarding the semiotics of sound, I have been able to conduct an in-depth analysis of the way sound functions in audio-described texts and have illustrated how the complex terminology regarding the functioning of sound in audiovisual texts can be turned into an effective analytical instrument. However, the method also has its limitations as a research tool. The method is less appropriate for an analysis of macro-structural features such as rhythm and identity chains, which ideally require a greater number of successive scenes to be analysed than the multimodal transcription method allows.

In addition, when applying the theoretical concepts developed in the literature review to a specific text, ambiguities can arise and it becomes apparent that some of the terminology used for annotation and transcription require fine-tuning. For example, my analysis underlined the importance of the audience's active information-linking of audio elements in the text when the correlation is not explicitly cued in the text. However, the category of information-linking is currently an umbrella term covering different types of information-link with varying degrees of explicitness. In addition, the difference between a sound with a confirmatory effect and one with an evocative effect seems to be a matter of degree (consider the example of the rain in *Loft*, for instance).

A more detailed subdivision of some of the concepts regarding sound—the variables that determine the degree of explicitness or the concept of information-linking, to give two examples—combined with more objective indicators for describers to assess the effects of sound is certainly required.

This observation is evidence of the great complexity of this field of study and the many variables that are involved in the meaning-making process of audiovisual texts. It explains why the development of a robust, exhaustive and systematic analytical model for analysing multimodal (translated) texts is and remains such a challenge.

References

Baldry, A. , & Thibault, P. J. (2006). Multimodal transcription and text analysis: A multimodal toolkit and coursebook with associated online course. London: Equinox.

Bordwell, D., & Thompson, K. (2013). *Film art: An introduction* (9th ed.). New York, NY: McGraw-Hill.

Braun, S. (2011). Creating coherence in audio description. *Meta: Translators' Journal, 56*(3), 645–662.

Chion, M. (1994). *Audio-vision: Sound on screen*. New York, NY: Columbia University Press.

Fryer, L. (2010). Audio description as audio drama: A practitioner's point of view. *Perspectives: Studies in Translatology, 8*(3), 205–213.

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's introduction to functional grammar*. London: Routledge.

Hirvonen, M., & Tiittula, L. (2010). *A method for analysing multimodal research material: Audio description in focus*. Electronic proceedings of the KäTu symposium on translation and interpreting studies. Retrieved from: https://www.sktl.fi/@Bin/40698/Hirvonen%26Tiittula_MikaEL2010.pdf (last accessed 05-10-2017).

Independent Television Commission (ITC). (2000). *ITC guidance on standards for audio description*. London: ITC.

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

Pérez-González, L. (2014). *Audiovisual translation: Theories, methods, issues*. London: Routledge.

Remael, A. (2012). For the use of sound. Audio description and sound: A few key issues. *Monti: Multidisciplinary in Audiovisual Translation, 4*, 255–276.

Remael, A., & Reviers, N. (2018). Multimodal cohesion in accessible films: A first inventory. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 260–280). London: Routledge.

Reviers, N., Remael, A., & Daelemans, W. (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). Bern: Peter Lang.

Reviers, N. (2016). On context and intersemiotic cohesion in audio description. In M. Kažimír (Ed.), *Kontexty: Interdisciplinárny zborník* (pp. 272–297). Gorlice: Ośrodek Kultury Prawosławnej.

Royce, T. (2007). Intersemiotic complementarity: A framework for multimodal discourse analysis. In T. B. Royce W. (Ed.), *New directions in the analysis of multimodal discourse* (pp. 63–109). Mahwah, NJ: Lawrence Erlbaum.

Taylor, C. (2003). Multimodal transcription in the analysis, translation and subtitling of Italian films. *The Translator, 9*, 191–205.

Tseng, C.-I. (2013). Cohesion in film: Tracking film elements. Basingstoke: Palgrave Macmillan.

Van Leeuwen, T. (2005). *Introducing social semiotics.* London: Routledge.

Van Looy, E. (2008). *Loft* [film]. Belgium.

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

---

1 The terminological and conceptional difference between the terms “audiovisual” and “multimodal” is not always clear and the terms often seem to be used interchangeably in the literature even though they represent different approaches. The present paper has taken a pragmatic approach and tries to remain true to the way these terms are used in the works cited. We therefore consistently refer to “audiovisual texts” and use the term “multimodal” when referring to aspects of meaning-making and cohesion.

2 The PhD project entitled “Audio Description in Dutch: A corpus-based study into the linguistic features of a new, multimodal text type” was funded by the BOF research fund of the University of Antwerp, and conducted at the Department of Translators and Interpreters under the supervision of Prof. Dr Aline Remael and Prof. Dr Reinhild Vandekerckhove, between 2012 and 2017. The thesis is available digitally in the library of the University of Antwerp.