MeMAD Deliverable

T5.1 Multimodal Annotation of Described video

Grant Agreement number	780069
Action Acronym	MeMAD
Action Title	Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy
Funding Scheme	H2020-ICT-2016-2017/H2020-ICT-2017-1
Version date of the Annex I against which the assessment will be made	3.10.2017
Start date of the project	1 January 2018
Due date of the deliverable	31 December 2018
Actual date of submission	31 December 2018
Lead beneficiary for the deliverable	University of Surrey
Dissemination level of the deliverable	Public

Action coordinator's scientific representative
Prof. Mikko Kurimo
AALTO –KORKEAKOULUSÄÄTIÖ, Aalto University School of Electrical Engineering,
Department of Signal Processing and Acoustics
mikko.kurimo@aalto.fi

Authors in alphabetical order		
Name	Beneficiary	e-mail
Sabine Braun	University of Surrey	s.braun@surrey.ac.uk
Jorma Laaksonen	Aalto University	jorma.laaksonen@aalto.fi
Tiina Lindh-Knuutila	Lingsoft Language Services	tiina.lindh-knuutila@lingsoft.fi
Kim Starr	University of Surrey	k.starr@surrey.ac.uk
Liisa Tiittula	University of Helsinki	liisa.tiittula@helsinki.fi

Abstract
This Deliverable is part of the MeMAD project's WP5, *Human processing in multimodal content description*, which explores human approaches to processing and describing audiovisual broadcast and media content (as a specific type of multimodal content), and compares them with machine-based approaches. In light of the advances and current limitations of machine-based approaches, and in line with the project's aim to advance this field, especially with regard to video scene description and audiovisual storytelling, it was decided that one of the project's work streams would focus on comparing machined-based and human methods for describing audiovisual content with the aim of identifying characteristic patterns of each method and informing the further development of machine-based algorithms. This Deliverable describes the work carried out in Task 5.1, 1 *Multimodal annotation of described video*, which was aimed at preparing the comparative analysis. The Deliverable first contextualises the work by referencing different types of human audiovisual content description and considering their advantages and drawbacks in the context of MeMAD (section 1). This is followed by an overview of current insight into human understanding and description of multimodal/audiovisual content, based on cognitive, pragmatic and narratological frameworks of human discourse processing and storytelling (section 2) and an overview of the current state of machine-based description of (audio)visual content and storytelling (section 3). The research design for the comparative study, which was developed in view of the insights described in section 2 and 3, is described in the remainder of this Deliverable. Section 4 focuses on the design of the audiovisual data corpus which we have compiled in this WP and the approach we have taken to its annotation; the final sub-section (4.5) outlines our approach to the data analysis and comparison, which will form part of the subsequent tasks in this WP.

Content

1 Introduction

1.1 Aims and rationale of this Deliverable

One of the objectives of MeMAD is to create the basis for an automatically or semi-automatically functioning model of multimodal content description, which can be applied to different contexts of use, especially the retrieval of content from broadcasting archives and the description of content for the benefit of sight-impaired people. By focusing on the description of video scene description and audiovisual content, the project aims to go beyond the state-of-the-art of automatic multimodal content description, which is currently mainly provided for still images such as photographs.

The specific approach that the MeMAD project has adopted is to combine advances in computer vision and machine learning with insights into human processing of multimodal content. Accordingly, WP5, *Human processing in multimodal content description*, aims to: a) advance our current understanding of the main principles, techniques and strategies of human-made video scene description by synthesising insights from previous research into human multimodal content description; b) use this understanding to identify differences and commonalities of human and machine-based multimodal content description, and to evaluate both types of description; and c) develop a human-based model of video scene description that is applicable to various usage situations. The short-term objective is to generate knowledge that can be used in the MeMAD project to inform the automatic analysis model. In the long term, WP5's findings can be used elsewhere in developing content description services and technologies.

The present Deliverable focuses on the first task in WP 5, i.e. Task 5.1 *Multimodal annotation of described video*, and documents the development of the multimodally annotated dataset of described video. This will serve as the basis for the comparative investigation into human and machine-generated descriptions.

The type of multimodal content description that is most relevant for the work in WP5 is **Audio Description (AD)**, which was originally conceived for the benefit of visually impaired people. AD makes visual imagery, audiovisual content and multimodal performances accessible for blind and partially sighted audiences by supplying a 'translation' of visual images – and also sound effects that are difficult to grasp without visual context - into verbal language. In the case of films and TV programmes, the verbal descriptions are first scripted, and then voiced and inserted into hiatuses in the audio track. They are designed to complement the other elements of the audio track, i.e. film dialogue, narration and/or major sound effects. Whilst the complementarity entails that AD is selective, AD is the most

elaborate type of visual content description that is currently available. However, AD draws heavily on human resource and is therefore expensive to produce.

A further type of visual description that can be identified is **content description for broadcasting archives**. This type of description is created to varying levels of detail, ranging from keywords to more elaborate descriptions of what an image or visual scene depicts. The main driver for producing content descriptions is the likelihood of the re-use of the content, i.e. the re-insertion of the content into another programme, in the future. Broadcasters therefore prioritise the description of content for which they own or have cleared or established the rights, i.e. which they can re-use internally or sell to other media companies. Content descriptions for archival purposes are used in written form only, as an ancillary text to the multimodal content, obviating the need for the descriptions to fit in audio hiatuses. Content descriptions therefore have the potential to be more comprehensive than audio descriptions, but—as in the case of AD—their creation requires resources. However, content descriptions tend to be more 'literal' or factual than AD, especially AD for filmic drama and movies, which can at times be 'narrative' or figurative.[1] A model of machine-generated content description is therefore likely to be a more achievable goal within the MeMAD project lifetime than a model for generating elaborate audio descriptions.

Audio description for visually impaired people – surrogate text; provides media access	**Content descriptions for broadcasting archives – ancillary text; retrieval aid**
• Scripted and then voiced and inserted into hiatuses in audio track so as not to overlap with the audio track	• Scripted and time-aligned, used in written form; no problems of overlap with the audio track
• High demands for coherence with other elements in the audio track (e.g. dialogue) due to shared use of audio track	• Lower demand for coherence with audio track, due to independent use of descriptions
• Time/space restrictions entail incomplete-ness, but complementarity and human ability to infer 'missing' information mitigate against information loss	• Fewer space/time restrictions facilitate a higher level for completeness where required, due to stand-alone use of the descriptions
• Less factual/literal, i.e. narrative rather than descriptive	• More factual/literal, i.e. descriptive rather than narrative

Table 1.1: Key features of different types of visual content description

[1] In The Hours (2003), for example, a father whose demeanour when looking at his young son at the breakfast table may indicate that he is anxious for his son to finish his breakfast is described as "point[ing] his finger in a mind-you-eat-your-breakfast kind of way at the boy", which includes an element of interpretation on the audio describer's part. In Avatar (2010), one of the exotic plants found on the mysterious Pandora planet is described as a plant with "spirals like concentric upside-down parasols", providing an analogy, which aids comprehension but is interpretive.

Given their characteristics, the content descriptions used by some broadcasters would be a good candidate for creating automated models of multimodal content description. However, these descriptions are an internal resource to broadcasters following internal rules of prioritisation. They are therefore not as widely and systematically accessible for research purposes as AD, which is increasingly available due to changes in broadcasting legislation such as the European Audiovisual Media Services Directive, 2010. Although the availability of AD varies in quantity, depth/detail and quality between countries and audiovisual genres (see further in section 4.2 Material selection), human AD continues to be the most elaborate type of (audio)visual content description in the public domain. It is a rich source of information about the visual elements in audiovisual content and—as will be explained further in section 1.2—a rich and relatively well studied source of insight into both how human understanding and human description of audiovisual content works. On balance, it is therefore a suitable basis for modelling multimodal comprehension and description, which the MeMAD project has hypothesised can inform the development of machine-based approaches to the description of audiovisual content. Based on this, WP5 focuses on the analysis of human AD and its comparison with machine-generated content descriptions.

1.2 Overview of the study of human audio description

With AD being one of the main objects of study in WP5, this section will give a brief overview of key current insights into AD to date. As was explained above, AD is a means to make audiovisual content accessible for visually impaired audiences by inserting short descriptions of the visual elements into hiatuses in the audio track. Audio descriptions are not intended to be stand-alone texts. They are created (by the audio describer) and processed or understood (by the blind audiences) in conjunction with those elements of the audiovisual content that remain accessible for visually impaired audiences, i.e. the dialogue and narration as well as many sound effects, music and song lyrics.

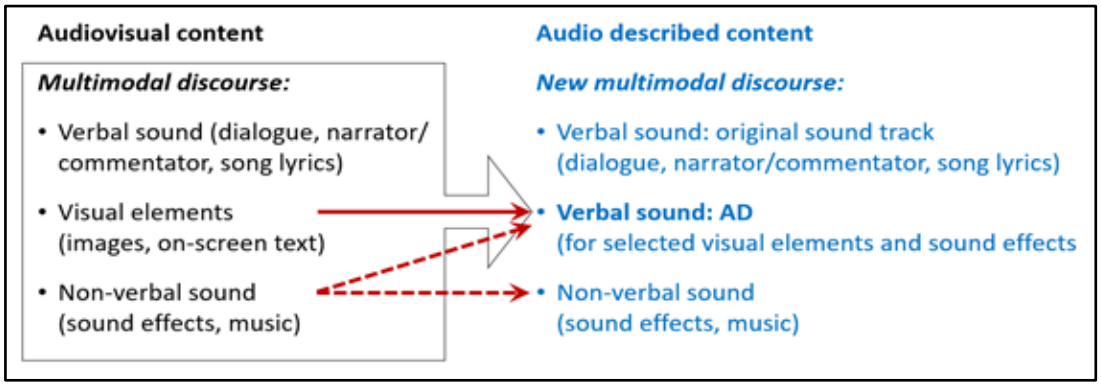


Figure 1.1: Audio description of audiovisual content as multi-/intermodal translation

The study of (human) AD is mostly situated within the field of Translation Studies, where AD is conceptualised as a type of multi-/intermodal translation as shown in Fig. 1, and more specifically as a practice of translating visual images or visual elements of audiovisual material, as well as sound effects that cannot be identified without seeing the associated visual elements, into verbal descriptions. In order to make sense of the audiovisual source, audio describers use their human ability to combine the different elements of the audiovisual narrative (e.g. visuals, dialogue, sound effects, music) into a coherent story in their mind (Braun, 2011, 2016; Vandaele 2012). Then they decide which of the visual elements and non-identifiable sound effects are crucial for understanding the story, before verbalising the information conveyed by these elements with the aim of enabling blind audiences to create a similarly coherent story.

Research has highlighted that the richness of the visual mode and the time restrictions imposed by the need for the audio descriptions to fit into gaps in the sound track require a complex approach to information selection and verbalisation on the part of the audio describer, involving strategies for the prioritisation of information (Fresno et al., 2016), and the use of strategies that ensure an optimal 'interaction' of the AD with the verbal and non-verbal sound (Braun 2011; see section 2.3 below).

Some AD research has focused on the linguistic realisation of different aspects of the description (e.g. Jimenez Hurtado, 2007 and Salway, 2007 on lexical choices; Zabrocka & Jankoswka, 2016 on co-speech gestures; Hirvonen, 2012 and Hirvonen & Tiittula, 2012 on visual and verbal representations of space in audio description; Hirvonen, 2013a on linguistic perspectivation strategies for filmic point of view; Hirvonen 2013b on similarities between visual and linguistic representations).

However, an area that has received attention more recently is AD style. In contrast to the classic position of AD practice, which was to 'describe just what you see', research suggests that narrative approaches which focus more holistically on the story that is told are more effective (Bardini, 2016, 2017; Kruger, 2010; Mälzer-Semlinger, 2012; Ramos Caro 2016). A case in point in this discussion are emotions, where the traditional practice is to avoid describing emotions while reception research has shown positive effects of conveying emotions in AD (Bardini, 2017; Ramos Caro, 2016).

Other research has focused on adapting the AD style and strategies to a given audiovisual genre (Davila-Montess & Orero, 2016 for adverts; Udo & Fels, 2006 for theatre; Orero, 2011 for children; Mangiron & Zhang, 2016 and Walczak, & Fryer 2017 for games and other virtual reality environments). In addition, research has highlighted the possibility of repurposing AD

to assist cognitively diverse audiences with emotion recognition difficulties (Starr, 2018) as a way of personalising audiovisual content.

An ongoing conundrum in AD research is the 'objectivity' debate, i.e. the question to what extent AD can and should be objective, given that visual content often leaves much room for interpretation, with meaning being in the eye of the beholder (Freyer, 2016; Mazur & Chmiel 2012; Remael et al. 2014). Whilst this debate has at times taken a prescriptive stance, arguing that subjectivity in AD should be minimised, several points in relation to this debate are noteworthy. On one hand, audiences often demand 'objective' AD (e.g. Lopez et al., 2018). However, they may know little about how AD is created and about the difficulties associated with achieving objectivity. AD is part of the post-production process; the audio describer does not normally have access to the director and his/her artistic intentions. Like any other recipient, the describer is therefore left to his/her interpretation of the audiovisual content in question. As will be expounded in section 2, the nature of human meaning-making entails that a degree of subjectivity in the interpretation process is inevitable in AD and other mediation/transfer activities. Thus, there will always be more than one acceptable solution, and this is further exacerbated by the fact that blindness is not a homogeneous condition and that there is a demand for personalisation of audiovisual content within this community. On the other hand, recent studies suggest that an element of subjectivity may, in fact, be desirable, for example, to convey emotions (Ramos Caro, 2016), to aid character identification (Wilken & Kruger, 2016), and to increase the recipients' feeling of 'presence' or immersion (Walczak & Fryer, 2017).

This section has given a brief overview of current insights into AD, including the complexity of information selection, prioritisation and verbalisation strategies; the advantages and drawbacks of different description styles; and the premise that whilst AD cannot be entirely objective, a degree of interpretation and subjectivity may lead to more successful AD. Perhaps with the exception of the desirability of some degree of subjectivity, these insights apply to content description for retrieval purposes as well. Given the relatively low level of sophistication that machine-generated descriptions of audiovisual content can currently achieve, the key characteristics of human multimodal content description are likely to create challenges for machine-generated descriptions. However, the MeMAD project aims to achieve significant progress in the automated description of audiovisual content. This makes it necessary to tackle these challenges. Arguably, an important prerequisite for this is to understand in more detail how human-meaning making works. This will be the focus on section 2.

2 Human understanding of multimodal content

2.1 Cognitive-pragmatic frameworks of human discourse modelling and storytelling

Among the plethora of theoretical models developed to study human communication, cognitive and pragmatic models of discourse processing offer great potential in the context of the MeMAD project, as they focus on explaining how we process and understand verbal and multimodal content and retrieve the messages or 'stories' (used here in a broad sense) communicated through it. Their particular strength is that they have developed plausible and powerful explanations of how human meaning-making works. These models take account of the human ability to infer meaning (Relevance Theory, Sperber & Wilson 1995) and to build mental representations of what is being communicated (Mental Model Theory, Johnson-Laird 1983, 2006), by virtue of combining perceptual input and prior knowledge. Other models focus on the role of background knowledge, thought to be stored as 'schemata' or 'scenarios' of places, activities or events, and activated through cues in a text or an image, as one of our main resources to make sense of what we hear and see (Cognitive Narratology, Herman 2002, 2013).

Cognitive-pragmatic frameworks have traditionally focused on mono-modal and mono-lingual communication, but there is a growing body of research using these frameworks to investigate multimodal communication (e.g. Dicerto, 2016; Mubenga, 2009; O'Halloran et al., 2014), multimodal translation (see Deliverable 4.1), audiovisual translation (e.g. Braun 2016; Desilla 2012; Kovačič 1993; Martínez Sierra, 2010) and audio description (Braun 2007, 2011, 2016; Fresno, 2014; Vercauteren & Remael, 2014).

In order to understand the full potential of AD to provide insights into human processing of multimodal content, the present section will outline what cognitive-pragmatic approaches offer for conceptualising human meaning-making in monomodal and multimodal discourse, or stories. The focus will be on Mental Model Theory, which underpins cognitive models of discourse processing; Relevance Theory, which presents the most comprehensive pragmatic model of communication; and Cognitive Narratology, which provides insights into the temporal and spatial unfolding of stories, especially in genres that are pertinent to audiovisual/ broadcasting content. These theories have been developed separately but they have complementary strengths which can be combined to conceptualise how we process multimodal content and the discourse or stories arising from them.

Mental Model Theory (MMT)

MMT is essentially a theory of human reasoning (Johnson-Laird 1983, 2006). One of its basic postulates is that "when individuals understand discourse, or perceive the world, or imagine a state of affairs […] they construct mental models of the corresponding situations" (Bell & Johnson-Laird 1998: 27). Mental models represent possibilities of how things could be in a given situation. In the process of reasoning and understanding, we draw conclusions about the plausibility of different possibilities based on what we know.

MMT has been used to model (verbal) discourse processing, i.e. to explain how we create mental models of situations described in texts (Dijk & Kintsch 1982, Brown & Yule 1983, Herman 2002). The beginning of a story (news item, text, novel etc.) normally gives rise to several possibilities, i.e. mental models, but as the story unfolds, we seem to settle on one of these in our interpretation of the textual cues (bottom-up processing) in light of our prior knowledge, the socio-cultural context of reception, and expectations raised by our prior knowledge, including schematic knowledge about places, activities and/or events (top-down processing). Mental modelling constitutes a process of hypothesis formation, confirmation and/or revision. The cues we bring to bear on this process vary in weight. A textual cue can be constitutive or decisive to our understanding, or it may confirm, reinforce, modify or contradict a previous understanding or hypothesis. Some cues are also redundant.

Through its focus on the different sources of cues for comprehension, MMT provides a useful starting point for analysing how we process discourse or tell and understand stories including in the context of audio description. Relevance Theory and Cognitive Narratology are complementary in that they elaborate on some of the details of this process.

Relevance Theory (RT)

RT provides a detailed account of how we understand individual and conjoined utterances in a text. It postulates that utterances are normally under-specified (e.g. by containing ambiguities that have to be resolved) and that as recipients we need to develop them into full-blown semantic representations (propositions) in order to derive the intended meaning (Sperber & Wilson 1995). According to RT, we achieve this by retrieving, as best as we can, the explicit and implicit assumptions (i.e. *explicatures* and *implicatures*) that a speaker is making. We normally begin by retrieving the explicatures. This involves working out the meaning of the key lexical items in an utterance (reference assignment), disambiguating words (e.g. pronouns) and pragmatically enriching what is said (e.g. working out causal, temporal and other links between utterances), resulting in a basic level of utterance understanding. The next step is then to retrieve one or several implicatures although these

steps can overlap (Wilson & Sperber, 2004), leading us to uncover a speaker's communicative message or intention.

RT claims that both explicating and implicating are highly inferential processes in which our 'cognitive environment', i.e. our knowledge and cultural experience, and the context we construe of the situation, play a significant role. RT asserts that these processes are guided by the human tendency to maximise relevance (*Cognitive Principle of Relevance*), which acts as a 'mechanism' that prevents us from infinite processing. As a consequence, RT argues, discourse processing is based on the assumption that speakers/storytellers normally wants to be understood and choose the optimally relevant way of communicating their intentions (*Communicative Principle of Relevance*). In accordance with this, we stop processing an utterance as soon as we derive an interpretation that we find sufficiently relevant. We are entitled to regard this interpretation as the optimally relevant interpretation as it provides the best balance between processing effort and effect. Utterances which require a high processing effort to reach this point (e.g. due to non-literal meaning) normally yield greater meaning effects. They are often richer in 'weak', i.e. more individual, implicatures.

This detailed account of how we work out utterance meaning 'step-by-step' highlights the human 'effort after meaning' (Bartlett 1932), i.e. our ability and perhaps conditioning to fill in unsaid details and supply links in the pursuit of making sense of someone's utterances and, more broadly, the world around us. However, to fully explain our ability to process *stories*, i.e. entire narratives, which normally have a beginning, a main part (problem and resolution) and an ending, it is useful to consider the main tenets of Cognitive Narratology as a complementary framework.

Cognitive Narratology (CN)

The emergent field of CN has been defined as "the study of mind-relevant aspects of storytelling practices" (Herman 2013). It builds on earlier models of Schema Theory, which postulate that our knowledge about the world—including knowledge about different types of events and situations—is organised through (stereotypical) schemata of these events or situations, which we derive from our experience (Bartlett 1932, Mandler 1984, Shank & Abelson 1977). Schemata are thought to be part of our cognitive system. They also include **narrative/story schemata**, i.e. abstract key elements of story structure that constitute knowledge about how different genres of stories are normally constructed. These schemata have become known as **story grammars** (Mandler & Johnson 1977, Mandler 1984, see also Appose & Karuppali, 1980).

Different story grammars have been developed to represent a basic story arc, and have sometimes been expressed in the form of rules, but in principle they include elements such as character(s) and setting; initiating event and initial response; plan, attempts or actions; consequence, outcome and resolution (Appose & Karrupali, 1980). Schemata are activated when we build a mental representation of a story (termed 'story world' by Herman 2002). They provide a 'skeleton' onto which cues from the story can be mapped. As a theoretical construct, they can explain how we derive complex interpretations of stories based on a small number of cues.

An important question for narratology is how we achieve **coherence** in narrative exposition, i.e. the impression of temporal and causal continuity of meaning and connectivity across the story arc. In a seminal work in text linguistics, Halliday and Hasan (1976) have analysed coherence from a semantic point of view, as a product of textual cohesion. Their model of text as a semantic unit that is 'bound together' by more than grammatical structure has led them to emphasise the role of lexico-grammatical cues on the text surface ('cohesive ties') in the creation of textual coherence. This approach has also been adopted in multimodality research, leading to a discussion of cross-modal links in multimodal texts in terms of 'intersemiotic cohesion' (e.g., Baumgarten 2008 and Chaume 2004 for films). However, continuing linguistic research has demonstrated that coherence is in fact a much more complex concept (e.g., Blakemore 1992; Beaugrande and Dressler, 1981; Brown and Yule, 1983; Bublitz & Lenk, 1999; Bublitz, Lenk *et al*. 1999; Gernsbacher and Givón, 1995) and has moved away "from reducing coherence to a product of (formally represented) cohesion and/or semantically established connectivity" (Bublitz 1999: 1) to a view that it is the text recipients who supply the links needed to create continuity of meaning and that formal cohesion is neither a necessary nor a sufficient condition for coherence.

Whilst this represents a shift from coherence as a semantic concept to coherence as a pragmatic, interpretive notion, a storyteller can select appropriate means of expression to support the creation of coherence in the recipient's mind by drawing on a comprehensive repertoire of linguistic resources, including, for example:

- Cohesive ties to make causal, temporal and other links explicit (Halliday & Hasan, 1976; Tanskanen 2006);
- Coreference chains to facilitate character and object identification and re-identification (Halliday & Hasan, 1976);
- Bridging inferences with typical exemplars to supply a range of semantic links (Myers et al., 2010);
- Motion verbs to create a sense of 'fictive motion' in a story (Talmy, 1983).

Furthermore, two types of coherence are normally distinguished, i.e. local coherence between adjacent utterances, which is supported by the use of the above linguistic resources, ang global coherence, which emerges from the overall topic and from consistency of e.g. style and register.

Focalisation (Bal and Lewin, 1983; Bal, 2009) as a function of both story and storyteller, creates an intermediate layer of narrative perspective (or 'bias') from which events are described and interpreted, suggesting that our understanding of story worlds is subject to influences which are not typically transparent or self-evident at first sight. For instance, our comprehension of events might differ greatly between the opposing narrative focalisations of victim or perpetrator of a crime, particularly where this kind of bias is revealed some way into a developing story arc. Human cognitive processing of narrative therefore requires engagement with issues of focalisation and bias in order to contextualise episodes of conflict and resolution.

2.2 Multimodal discourse modelling/storytelling through the cognitive-pragmatic lens

The previous section focused on verbal discourse, in line with how theoretical frameworks evolved. Given the focus of the MeMAD project on multimodal and specifically audiovisual content, the present section outlines how the frameworks introduced in section 2.1 above can be applied to understanding and conceptualising multimodal storytelling.

MMT claims that mental models can be created on the basis of visual perception as well as verbal discourse, emphasising that "[m]odels of the propositions expressed in language are rudimentary in comparison with perceptual models of the world, which contain much more information— many more referents, properties, and relations" (Johnson-Laird 2006: 234). Sperber and Wilson do not have much to say on visual or multimodal discourse, but from their claim that visual images as "non-propositional objects" do not have explicatures (1995: 57) and given the importance of explicatures in RT, the theory might appear less applicable to multimodal discourse. However, various suggestions have been made to adapt RT to the analysis of multimodal discourse, arguing that visual images may give rise to both explicatures and implicatures (e.g. Braun 2007; Forceville 2014; Yus 2008). CN has been applied to both monomodal and multimodal storytelling, especially in filmic narrative (Herman 2002).

One question to be answered is therefore how, according to these models, meaning arises from multimodal content, and specifically audiovisual content. The characteristics of the different modes of communication provide a useful starting point. As Kress (1998) notes,

verbally told or written stories unfold temporally and sequentially, while the visual mode presents information spatially and concurrently. The verbal mode explains, describes, narrates and classifies; visuals display and arrange elements in space. However, because audiovisual content normally also "sequentialises and temporalises visual images" (Kress 1998: 68), it can be said that meaning in audiovisual content essentially arises from visual-verbal co-narration; non-verbal audio such as sound effects and music further contribute to this. In the opening scene of *Notting Hill*, for example, a montage of Julia Roberts alias Anna Scott showing scenes of her glamorous life is pervaded by the music, rhythm and lyrics of the Aznavour song "She" to 'tell' us her story and introduce her as a superstar. Notably, the song's famous refrain (a drawn-out "she") coincides with close-ups of Anna's face as she smiles into the paparazzi's cameras or waves at the cheering crowds. In the next scene, the male protagonist, Hugh Grant alias bookshop owner William Thacker, speaks in his own voice as he is walking us through Notting Hill to introduce us verbally and visually to his more ordinary life, friends and neighbourhood.

As Lemke (2006) asserts, when different modes of expression are combined, their meanings are not simply added to each other; they contextualise, specify and modify each other. Thus Anna Scott is not simply *identified* in the opening scene. The explicatures and implicatures that we derive from the song lyrics, the cheers of fans, the flash photography, the close-ups of Anna's face and her appearance on the covers of glossy magazines (e.g. 'people on glossy magazines are famous' as a simple implied premise) create a mental model that *glorifies* her, whilst the inferences encouraged by William's casual tour of Notting Hill, supported by the expectations arising from the genre of romantic comedy, suggest that he is an 'ordinary guy'.

Johnson-Laird (2006: 233) maintains that the cognitive processes involved in integrating cues from different sources into mental models are not well understood yet. Arnold & Whitney (2005: 340) believe that we have dynamic strategies for "weigh[ing] all the available cues according to their relative reliability". The stages of explicating and implicating assumed in RT provide a basis for elaborating on this, but the crucial point here is that cognitive-pragmatic frameworks of discourse processing highlight the important role of the recipient's cognitive environment (see RT) in identifying and interpreting the cues from different modes and the cross-modal relationships that contribute to meaning-making in multimodal discourse. Many of the explicatures and implicatures arising from the introduction of Anna Scott in *Notting Hill* will be based on fairly universal knowledge about superstars. Most viewers will also be able to create meaning from William's comment that Notting Hill has street markets "selling every fruit and vegetable known to man". By contrast, knowledge about the district's evolution into a trendy part of London may be less widely available, but where it is, it could aid the interpretation of the visual snapshots of Notting Hill and add

detail to modelling William's character. Differences in the recipients' cognitive environments will thus lead to intersubjective differences in story interpretation. Equally important, these differences are likely to be magnified when visual images are involved, as visual meanings are "construed largely as a result of tacit learning", making them "more open to idiosyncratic interpretations" (Jamieson 2007: 34) or, in RT terminology, 'weak implicatures'.

The theoretical considerations of human multimodal discourse processing/storytelling make it clear that this is a complex process with a range of uncertainties; they explain why we draw different conclusions from the same premises and why storytelling may be unsuccessful. Whilst by emphasising the subjectivity of discourse/story interpretation, these models allude to the potential for creativity (which can, for example, be exploited in making sense of art works), the complexity and subjectivity of human discourse modelling also means that it has to date largely eschewed systematisation and formalisation. Bearing in mind the aims of WP 5 of the MeMAD project, the next section will consider some of the implications of what we know about human discourse modelling/storytelling for multimodal content description, (with specific reference to audiovisual content).

2.3 Multimodal content description through the cognitive-pragmatic lens

As was outlined in section 1.2, there are currently two main types of multimodal content description, i.e. audio description for visually impaired audiences and content description for archival purposes. For the purpose of analysing human descriptions and comparing it with machine-generated descriptions in WP5, the focus will be on AD, for the reasons given in section 1.2

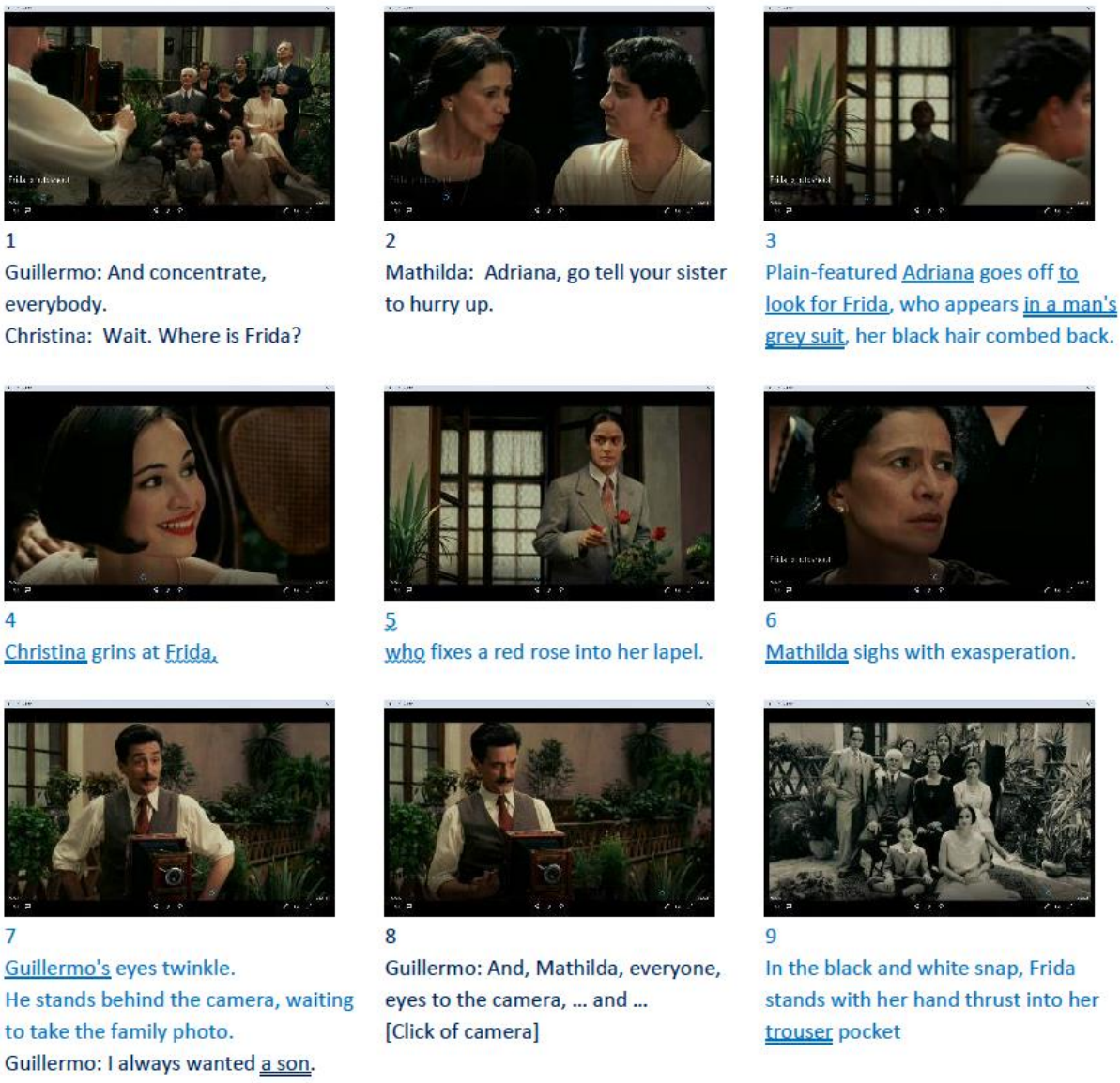
To reiterate, human audio describers normally need to be highly selective with regard to the visual information they describe and the amount of detail they can include, because the descriptions need to fit into gaps in the audio track. AD has therefore sometimes been characterised as 'partial' translation (Benecke 2014). However, the cognitive-pragmatic frameworks outlined above can be used to explain why such labels do not fully do justice to the complex processes of comprehension and (re-)production of meaning that are associated with AD. Whilst the textual surface of AD will only provide a partial representation of the visuals included in the audiovisual source material, the human ability to draw inferences, build mental models and create coherence by combining cues from the AD text with information in the audio track, their world knowledge and expectations ensures that information which is not explicitly included in the AD, i.e. may appear as omissions, can still be retrieved by the recipients. Considering the differences between the visual and the verbal mode of expression, especially the sequentiality of the verbal mode, which means

that 'telling' takes more time than 'showing', and the richness of the visual mode (see also below), some degree of selectivity will arguably be required for any type of multimodal content description.

The important point is that audio describers **need to identify the most relevant and narratively salient cues** to render narrative force (Vandaele 2012). RT makes it clear that this process is not simply aimed at 'filtering' out irrelevant information. According to RT, we are encouraged to believe that all elements of a story are optimally relevant. (The possible conclusion that an element is not relevant is, in fact, a less desirable result, meaning that the story is not entirely successful.) Rather, the process of identifying the most relevant and salient cues involves an in-depth analysis of the visual-verbal relationships (Dicerto 2016) and an assessment of different 'translation' strategies in light of audience requirements.

Whilst Diaz Cintas and Remael (2007: 49) bemoan the fact that audiovisual translators "will never have enough time to carry out an in-depth script analysis", cognitive-pragmatic frameworks emphasise the importance of exactly such an analysis and more – a multimodal analysis. Conclusions about the relevance of a particular story element will not normally emerge until the completion of this analysis and can only be drawn in relation to the chosen 'translation' strategy.

In the extract from *Frida* (2002) below, for example, the AD generally centres on the characters in focus (Adriana, Christina, Frida, Mathilda, Guillermo). Most characters are only named, with their main action described in a grammatically simple sentence ("Christina grins at Frida", "Mathilda sighs with exasperation", "Guillermo's eyes twinkle"), but Adriana and Frida are assigned brief descriptions of their appearances ("plain featured" and "in a man's grey suit" respectively). The next dialogue turn (G: "I always wanted a son.") makes it clear why the detail of Frida's appearance is crucial for coherence. The final AD section reinforces this message by referring to Frida's trouser pocket. The audio describer's decision to include a description of Adriana's appearance is less obvious, but the description ("plain featured") contrasts with the glamorous appearance of Frida, marking Frida as the central character of the story. The other women's reactions to Frida's lateness and grand entrance ("Christina grins", "Mathilda sighs with exasperation") reinforce this and add to the narrative power.



Example 2.1: Frida – taking a family photograph, old style

At the same time the above example also illustrates that audio describers **add aspects that are not directly visible and only inferable.** This relates to Gutt's (2000) observation that translation involves not only identifying the explicatures and implicatures in the source discourse but replacing and/or 'redistributing' them in the target discourse to provide for differences in the source and target recipients' cognitive environments. Here, for example, an assumption that is implicit in the visual narrative, namely that Adriana goes off to look *for Frida* (1) is made explicit in the verbal description. There is no visual element that provides the reason why Adriana walks away; we infer the reason from the preceding dialogue turn. Similarly, the assumption that Christina's grinning is directed *at Frida* (4) is only inferable from the direction of Christina's gaze, from our understanding of the preceding and

subsequent shots (3 and 5), in which Frida arrives on the scene, as Christina and Frida are not shown together in shot 4.

Furthermore, the **richness of visual images** raises the question of the most efficient way of describing, i.e. whether it is more efficient to state the explicatures arising from the images, leaving it to the audience to derive appropriate implicatures, or whether the description should verbalise the implicatures to save time. In the example below, taken from the opening scenes of *The Hours* (2003), the AD relating to 1-3 spells out some of the explicatures first, by taking us through the physical details of the woman's attempt to fasten the buttons and belt of her coat (note that we do not see the woman in full) while leaving us to infer that she is getting ready to go out. By contrast, the AD relating to 4-5 focuses on a simple implicature from the images, namely that the woman is sitting down and is writing something. The further-reaching implicature, that she may be writing a suicide note, is not spelt out as the audience may retrieve this from the narrator's voice that is reading the content of what she is writing. There are, however, further visual cues in this scene which reinforce the suicide note hypothesis (e.g. the note is shown being left on the mantelpiece as the woman walks to a nearby river and begins to puts small rocks in her coat pockets). All of these cues are selected for description, in line with the goal of the AD, this being to create a coherent story.

1 2 3

A woman's slender hands tremble as she fastens the buttons and ties the belt of her tweed coat.

4 5

Earlier she sits writing.

Example 2.2: The Hours: Describing at different levels

In conclusion, the complexity of human processing of multimedia content means that capturing and systematising the essential characteristics of *human descriptions* for multimodal content is a highly complex task. The complexity of the processes involved in deriving good and meaningful descriptions of audiovisual content may also serve to explain current limitations in the efforts to automate such descriptions. At the same time, the prospect that different styles of description and different levels of granularity may return useful descriptions, by exploiting human inferencing and mental modelling powers, may mitigate against some of the current problems with producing elaborate video scene descriptions, for instance the over-use of generic vocabulary, lack of continuity and linkage between individual shots/images and so forth. In other words, existing machine-generated descriptions will at least provide a starting point for an analysis that can identify recurrent patterns of problems and thus highlight where the main issues arise. This will generate insights into how their potential for meaning-making can be improved.

The current state of the art of computer-generation machine description and visual storytelling will be outlined in section 3 below. The system of annotation that we have developed for the comparison of human and machine-generated content descriptions (section 4) is agile to accommodate the anticipated evolution of the descriptions during the life of the project.

3 Computer Vision, Visual Storytelling and Machine Description

3.1 Introduction to machine learning for visual storytelling

Until recently, automatic multimodal content description has consisted of techniques that detect visual and auditory elements from multimedia, and label them with pre-defined keywords or indexing concepts. Such keywords can be words derived from visual and aural categories and/or words recognized with a speech recognizer from the spoken utterances. This approach has severe limitations as, for example, accurate description of actions and properties of the visible objects has not been possible because the existing sets of labelled training data, on which all methods of automatic image recognition rely, have focused more on nouns as object classes and less on adjectives and verbs.

As a very recent trend, large image and video corpora, such as Microsoft Research's COCO (Lin et al., 2015) and MSR-VTT (Xu et al., 2016), respectively, have emerged. These datasets contain multiple human-written full sentence annotations (captions) in unrestricted natural English language for each image or video object. Moreover, some image databases, such as the *Visual Genome* (Krishna et al., 2017), provide both sentence-based and scene-graph-based annotations. In the latter case, the natural language annotations can be localized to

specific parts of the images, to describe just some details of the whole view. These developments in the availability of training and testing data have opened up new avenues for devising more accurate and efficient methods for automatic multimodal media data description.

Furthermore, deep neural networks have been found to provide superior performance in many visual machine learning and media analysis tasks. The success stories of deep neural methods include visual feature extraction and classification, and the implementation of recurrent encoder-decoder language models for translation from the visual domain to natural language. The modern approach to automatic image and video captioning is based on using deep convolutional neural networks for feature extraction or visual input encoding (Krizhevsky et al., 2012, Szegedy et al., 2014, He et al., 2016). This representation is then fed to a recurrent neural network, typically a Long Short-Term Memory (LSTM) network (Hochreiter et al., 1997), that decodes this visual encoding to an output sequence of words, a sentence or a caption that describes the multimodal content.

Training the word sequence decoders for image and video content description has conventionally been based on minimizing the cross entropy between the sentence generated by the model and the desired output. This approach is generally well-motivated theoretically, but does not aim to directly maximize any automatic performance measure used in practice such as BLEU, METEOR or CIDEr scores. In order to improve the captioning performance with respect to such automatic measures, researchers have started to use reinforcement learning (Ren et al., 2017) in training the captioning models. This has lead to clearly better results when measured by the automatically obtainable scores. Despite the significant recent progress, the current image and video description techniques are, however, still very unreliable, producing different textual descriptions for visually very similar contents.

3.2 Computer vision datasets for media captioning research

The most important computer vision datasets available for media captioning research are listed and characterized in the following table:

name	content	# objects	# captions	reference
Flickr30k	images	31783	158925	(Plummer et al., 2016)

MS-COCO	images	123287	616767	(Lin et al., 2015)
Conceptual Captions	images	3178371	3178371	(Sharma et al., 2018)
VisualGenome	Images + graphs	108249	5408689	(Krishna et al., 2017)
VIST	Image sequences	20080	100400	(Huang et al., 2016)
TGIF	video w/o audio	125713	125713	(Li et al., 2016)
MSVD	video	1969	80800	(Chen et al., 2011)
LSMDC	video	108536	108536	(Rohrbach et al., 2015)
MSR-VTT	video	6513	130260	(Xu et al., 2016)

All the above datasets are in open access and they have already been used or will be used in the experiments of MeMAD Work Package 2.

3.3 Image Sequencing

As a step beyond the automation of descriptions of individual visual images, the automation of *sequenced descriptions within a static image* environment (Huang et al., 2016; Smilevski, 2018) has developed apace, most notably in relation to the description of object inter-relatedness within single frame images (Krishna, 2017). Meanwhile, progress in machine-generated *descriptions for moving image sequences* has moved at a more modest speed (Xu et al., 2016; Rohrbach et al., 2017) due, in large part, to the dearth of sufficiently sizeable training and test datasets required to assist machine learning. Nevertheless, a range of innovative approaches have been trialled: the exploitation of temporal structures (Yao et al., 2015), question-answer techniques (Wu et al., 2016), video-sentence pairing (Venugopalan et al., 2015) and visual attention strategies (Xu et al., 2015; Kim et al., 2018). Regardless of whether the data adopted for the purposes of training computer vision models comprise still or moving imagery, however, the holy grail remains to produce a model for creating machine-generated, intuitive and coherent storytelling across multiple images read in sequence.

Fundamentally though, sequences of still images and (continuously) 'moving' images, i.e. video scenes, embody the same properties and may, superficially at least, be regarded as

posing the same challenges in terms of automating descriptions. Firstly, while short sequences of images frequently contain persons or objects that recur across the piece, and should therefore be regarded as prime candidates for conveying information of narrative saliency (see also Example 1 (from *Frida*) in section 2.3 above), variations in scale or placement may currently confound the automatic identification of continuity cues. Initially, this impacts the identification of key protagonists and action-relevant objects, subsequently inducing a knock-on effect where abstract concepts associated with these entities are also disregarded (e.g. failure to identify an image as relating to a group of 'friends' may also impact the visual-semantic association that cross-references a social gathering). Secondly, backward- and forward-referencing of objects and concepts between connected images ('inferential bridging') is still in its infancy, and consequently a consistent means of establishing coherence between frames within sequential moving imagery remains, as yet, largely out of reach. Although moving images and sequences of still images have similarities, they represent different challenges in this respect, as film imagery generally depicts composite motion sequences at a more granular level (specifically, 25 images per second) than would be expected from a sequence of five or six related still images from a Flickr album (Huang, 2016). Action identification and coherence should theoretically be more attainable in the former, given the advantage of more dense visual information.

Issues of inter-relatedness between people and objects in sequential imagery, both moving and still, represent a major milestone in automating descriptions, with the 'who did what to whom' question (who is talking to whom?) still posing a significant challenge which remains unresolved. Hypothetically, the addition of audio cue isolation to the computer vision model should assist in the disambiguation process. One avenue worth exploring is whether audio event detection and speaker diarization could assist in the identification of characters and sound-associated objects. Audio events comprise audible data attributable to specific actions, including elements such as speech, non-verbal utterances, animal noises, vehicle sounds, doorbell and telephone rings, and so forth. Automatic classification of these sound artefacts is referred to as audio event detection (AED) and can be applied to a range of practical applications, such as speech and speaker recognition (Babaee et al., 2018). Current methods for achieving AED include audio "preprocessing, feature extraction and classification methods" (Babaee et al., 2018: 661). Within the spectrum of opportunities this affords is the determination of specific prosodic features, capturing pitch, volume and duration.

Automatic speaker diarization, on the other hand, "is the process of partitioning an input audio stream into homogeneous segments according to the speakers' identities" (Vallet, Essid & Carrive, 2013), promoting the identification of speech events and turn taking between individuals in a shared audio event (e.g. a talk show), such that each speaker's

entry and exit points are recorded (speech repartition) and data, including cumulative speaking times, is captured. Work combining speaker diarization with visual data cues, notably changes in camera shot which focus on the current speaker, have refined the concept of a correlation between those who are speaking and those who are featured in the visual content. This link extends to the automatic identification of persons featured across multiple frames. It is achieved not by means of facial feature recognition, as this sits outside the scope of current machine learning techniques in multi-camera, fast moving audiovisual material, but through the use of features contained in the speakers' clothing. In summary, there is an existing precedent for combining audio and visual features to produce basic indicators of speaker coherence across narrative.

Pairing automated audio event extraction and speaker diarization with image sequencing models, were this to prove feasible during the lifetime of the MeMAD project, should exponentially improve continuous character identification between frames, eased by the extraction of a speaker's combined vocal and visual 'DNA'. Audio tagging of principal characters would likewise mitigate computer vision confounds arising where abstruse camera angles or abrupt changes of scale impede the machine in identifying reoccurring characters (or audio-defined objects, such as a barking dog). Combining audio and visual cues to infer continuity would therefore contribute significantly to creating narrative coherence in automatic descriptions. If this approach proves tenable, we believe our human annotation and analytical methods, which will be outlined in the remainder of this Deliverable, are sufficiently agile to accommodate a comparative analysis between the combined sound-image machine-generated descriptions and their human-generated equivalents.

4 Methodological Approach: Research design, materials, data processing and annotation

4.1 Research design

In accordance with the original project proposal, task 5.1, *'Multimodal annotation of described video'* (M1-M12), has focused on four principal components:

(i) the construction of a corpus of audiovisual materials consisting of human audio descriptions and original film dialogue in at least one of the project languages;

(ii) identification of short extracts within the corpus which lend themselves to human vs. machine generated description comparisons;

(iii) annotation of this audiovisual content in a manner which facilitates a comparative study featuring human and computer-generated video description;

(iv) preliminary analysis of parallel datasets (human annotations, AD and pre-development experimental machine-generated video descriptions) to pilot the methodological design and initiate first improvements in automated descriptions.

Each item is a key step in the preparation for comparative analyses between human-generated and machine-generated video descriptions in later tasks, and has required experimental modelling and piloting in the past year, in order to establish the optimum approach for the purposes of the project. Looking ahead to future tasks, it is anticipated that the results of our human analysis will be used to model human AD in T5.2 (M4-M18)/5.3 (M13-30), informing the development of machine descriptions in T2.2/2.3, and producing the 'Best Practice Guide for Video Descriptions' required for deliverable D5.3.

4.2 Materials

Selection of Materials

While audio described content is more readily available than other types of multimodal content description (section 1.1), being used by some broadcasters and content producers to enhance accessibility for sight-impaired audiences, the sourcing of audio described broadcast and digital media content is not without challenges, regardless of host territory. Many countries fail to offer AD for sight-impaired audiences, while others (the UK being an example) are moving towards a level of described content which exceeds statutory requirements (close to 20% of all broadcast programmes in the case of the BBC and Channel 4). The availability and quantity of audio described content varies widely according to the legislative frameworks in operation in each country with many territories remaining unregulated, despite moves by EU legislators to encourage wider participation and equal access to broadcast media for citizens (Council Directive 2010/13/EC, 2010). Furthermore, the economic viability of supporting additional post-production costs within the programme-making process represents a considerable burden for television producers such that AD, even where present, may be minimal or of variable quality and quantity. One area where audio description services are becoming noticeably more available is the streamed content sector. Netflix and Amazon Prime both offer growing catalogues of audio described original content on the UK sites, most notably in the high production-value drama and film genres, and this can also be offered in a range of languages (Spanish, German, Italian, Russian, Hindi etc.). Some film and television-derived DVDs also contain audio descriptive tracks, although the number of productions available is limited by the late arrival of AD to the industry (1990s onwards) and the current audience shift towards digital and streamed movie platforms,

which has impacted the number of contemporary DVDs produced with audio description tracks.

In addition to the availability of audio described materials generally, stylistic factors, both in terms of the density of audio insertions and their granularity in relation to the narratively salient details, means much current television production content is of limited use to the audio extraction processes originally envisaged within Task 5.1. An example of the type of issues encountered was highlighted during the pilot phase of this work package, when the serial drama genre was explored as a potential source of multimedia data for the purposes of investigating human vs. machine generated video descriptions. Episodes of *EastEnders,* a serial drama/'soap' produced by the BBC in the UK, were examined for quality and quantity of human audio descriptions. While this material contained useful examples of the kinds of narrative action which could theoretically inform human meaning-making in story-telling, the extent of the AD was constrained by quick-fire direction (multiple, very short scenes and rapid shot-changes) and a shortage of audio hiatuses. Hampered by these technical parameters, the corresponding AD was minimal, largely becoming a vehicle for announcing changes of location ("in the pub…") or for introducing new characters ("Bernadette and Tiffany arrive"). Documentaries, as an alternative genre of programming containing AD, also proved problematic. With the exception of flagship programmes such as the BBC's *Blue Planet*, where worldwide distribution rights positively impact production budgets, documentaries generally contain minimal AD, even in circumstances where the material naturally lends itself to colourful descriptions. Documentaries may also lack a clear narrative, with isolated segments failing to deliver 'intact', self-contained, micro-plots.

By contrast film productions, due to their long-form narrative exposition, lend themselves to more elaborate and narratively sophisticated storytelling and AD scripting, with opportunities for the describers to paint an audio picture which does more than merely label the characters and their locations (see sections 1.2 and 2.4). Poetic and evocative descriptions of cinematographic elements, as well as interpretive commentary on the narrative importance of key actions and events, elevate film AD from a mechanism for streaming basic information to a rich and colourful art form. This greater emphasis on explication in film storytelling is frequently matched by a richer lexicon and more complete descriptions than would be found in a standard television production. Our pilot study suggested these dual aspects, rich descriptions and contextualisation of content, distinguished feature film audio descriptions as the most comprehensive source of audiovisual data available for informing the creation of automated machine-generated descriptions. In theory, at least, film AD should facilitate visual information extraction, serving as a ready-made comparator for evaluating computer outputs.

However, while AD has a perceived value in the context of informing machine-generated video descriptions, our pilot stages also show that extracting *comprehensive* visual information from AD can prove problematic. AD embodies both the 'science of communication' and the 'art of omission', in the sense that inferential processes of meaning-making, mental modelling and coherence creation (the science), are played out through the audio describer's personal filter of individual interpretation, life experience and intuition (the art), all of which are tested against the benchmarks of redundancy and saliency. As a result AD is a highly personal production, drawing on the describer's interests, interpretation and individual biases, meaning that there is considerable potential for error and omission. Standardisation of AD has been a long time coming. Perhaps not unsurprisingly, however, given the complex nature of human meaning-making (as outlined in section 2), the application of rule-based methodologies for arriving at audio described outputs (Audetel/ITC, 2000; AENOR, 2005) has proved largely untenable, with a lack of consensus between describers about what should be included and omitted in a narratively complementary script (Vercauteren, 2007: 139; Yeung, 2007:241; Ibanez, 2010:144). The solution is most likely to be found in compromise and flexibility of approach, rather than dogma. However, this lack of standardisation naturally impacts objectivity, with considerable variation between describers in the way they choose to prioritise film material for inclusion in the AD, and the lexical breadth with which they choose to describe the selected elements (Matamala, 2018).

In addition to these constraints, the absence of suitable hiatuses in the audio track, due either to inopportune timing or a density of dialogue (or both), often shackle the describer, limiting the extent to which any supplementary visual information can be inserted into the source material. The result is that an 'internal negotiation' occurs between the audio describer's natural inclination to voice all relevant information, and the 'golden rule' of AD that prohibits interruptions to the original sound track (Hyks, 2005). As was highlighted in section 2.4, this is not such a sizeable problem for AD recipients, usually blind and partially-sighted audiences, as omissions in the AD will often be mitigated by the use of inferencing strategies, resulting in a more or less complete comprehension of narrative. Computer vision algorithms, on the other hand, currently lack complex inferential capacity which means that the AD alone cannot provide sufficient data to serve as a 'complete solution' for training machines to produce human-like descriptions. In summary, while it is unquestionably a useful source of visually descriptive information, closer inspection during the pilot stage has revealed that AD taken in isolation cannot offer a 'one-stop-shop' solution for informing the development of human-like machine-generated descriptions of moving images. A summary of key issues can be found below (*Table 4.1*):

Advantages of Film AD	Disadvantages of Film AD
focussed on visual imagery	not a complete narrative, but rather a 'constrained' supplementary text
may contain cues for key narrative events: characters, actions and locations	key narrative events may alternatively be relayed via other audio channels (dialogue, sound effects, original music score etc.)
can be lexically rich and eclectic	Choice of lexicon may be too sophisticated or subjective for direct comparison with machine descriptions
where sufficient hiatuses occur in the original audio, evocative descriptions can inform deeper immersion in film text	paucity of hiatuses in the original audio may limit the extent of, or preclude, AD
more reliable source of narrative cues than subtitles/dialogue alone	personal 'take' on plot interpretation and therefore not 'definitive'
subjectivity may be at the heart of 'human touch' AD	not objective

Table 4.1: Audio Description – Advantages and disadvantages for informing machine-generated descriptions

While these issues will be examined in some detail below, it cannot be overstated that – as highlighted in section 1 – AD for motion picture (movie) productions remains the most complete audio descriptive data resource available in respect of the visual content of moving images, and for this reason it is possible to make a compelling argument for using audio described films as a point of departure in defining resources applicable to task 5.1.

Audiovisual Corpus

Our primary experimental corpus, numbering fifty feature-length films, was drawn from a limited catalogue of audio described productions currently available on commercial release in DVD format through online retailers. Five movie genres, representing a diversity of cinematic styles, were chosen for analysis: comedy, action, thriller, 'romcom' and drama. Historical dramas containing anachronistic references, e.g. period costume, and animated productions featuring cartoon characters, were intentionally excluded in the knowledge that they were likely to confound computer vision applications which rely heavily on training data compiled from contemporary still and moving image datasets, paired with crowd-sourced captioning (*e.g.* the Microsoft COCO dataset, detailed in Lin et al., 2015).

Identifying 'Story Arcs'

Acknowledging the important role of story schemata in comprehension of multimodal discourse (section 2.1), our first step in data preparation was to identify a series of 'story arcs' within each feature film. These took the form of short stories-within-a-story (micro-narratives), containing clear, narratively significant beginning and end-points, and illustrated elements of crisis and resolution. Extracts were drawn from full-length feature films due to the availability of high quality audio description, however, it has not been the intention that they would be treated as part of a narrative with greater reach than the parameters of the extracts themselves.

Mindful of the lack of sophistication in current machine-generated video descriptions, we selected examples of basic social interaction as the focus for our data mining exercise. Uniform parameters were applied to the selection of 'story arcs' in order to standardize the dataset, and facilitate meaningful comparison and evaluation between human descriptions and those produced by machine learning techniques:

Category	Criteria	Observations
Source Text	Must contain audio description	Required to explore value of AD for informing computer-generated descriptions
Persons	1 or 2 principal characters	Incidental characters and small groups of people in the background of shots also permitted.
Actions	Minimum of 4 or 5 simple, common actions	e.g. sitting, running, talking, walking, hugging, kissing
Duration	20 secs – 3 minutes	Limited duration story arcs should simplify sequence modelling
Storyline	Self-contained micro-narrative	e.g. initiating action/crisis, proposed solution, action based on solution, consequence, result
Objects	Unlimited	Although no limitation was put on the number of objects in an extract, only those objects regarded as key to the action were included in our annotations

Table 4.2: Criteria for selecting 'story arc' extracts

Thus, in order to avoid a level of narrative complexity likely to defy current machine-generated description capabilities, scenes were selected on the basis that they contained one or two principal characters only, behaving or interacting in a naturalistic, socio-representational manner. Simple actions such as sitting, walking, talking, running, hugging and kissing occur frequently in film material (Salway, 2007) and for this reason are especially relevant to the improvement of simple, machine-generated video descriptions which currently fail to register these basic movements consistently and accurately.

While film presentations typically have a duration of between one and a half and two and half hours, the number of 'story arcs' available within each production varies according to narrative composition, directorial choices, and cinematographic presentation. For this reason, and in order to set an achievable goal, our target was to identify between ten and twenty 'story arcs' which met our selection criteria per film. We set a ceiling of twenty extracts per film in order to avoid over-representation by any one audio description style, production house or describer. This approach resulted in a corpus of approximately 500 extracts for annotation and analysis.

Story Arc: Example

Selected 'story arcs' take the form of short micro-narratives occurring within the context of a full feature-length film. Essentially, each 'story arc' represents both a dramatic episode salient to interpretation of the wider narrative, and a self-contained mini-plot in its own right. The duration of 'story arcs' was maintained between 20 seconds and 3 minutes in order to ease the application of sequence modelling techniques during later machine iterations.

An example of one such 'story arc' (Boy in a Field) is provided in *Figure 4.1* below, and is taken from the film *Little Miss Sunshine*. At the beginning of the extract a dispute arises between a teenage boy and family. The dispute is subsequently resolved by the intervention of a young family member. Screenshots of narratively key frames from the scene sit alongside a brief description of the action, provided in linear fashion:

	On a family road trip, a teenage boy (Duane) discovers he can no longer follow his dream of becoming a fighter pilot. He demands the camper van the family are travelling in is stopped, and he jumps out. Refusing words of comfort from his mother, he runs into an empty field, and sits down alone, to contemplate his future.

	Duane's young sister (Olive) offers to talk to him. She leaves the rest of the family back at the roadside and walks down a grassy slope towards her brother.
	Olive crouches down behind Duane, and without speaking …
	… puts an arm around him, leaning her head tenderly on his shoulder.
	Comforted by her presence and the knowledge that she truly understands his despair, Duane relinquishes his anger. They both rise …
	… and walk back towards the roadside where the rest of the family are waiting for them.

	In a sentimental, reciprocal declaration of affection, Duane resumes his role as 'big brother', carrying his little sister up the sharp incline near the road.

Table 4.3: Boy in a Field (Little Miss Sunshine)

In the above extract, we observe a typical film crisis-resolution scenario, in which the crisis (boy learns bad news) precipitates action (the boy leaves a parked van and sits alone in a field), followed by crisis resolution (his little sister comforts him), through consequences of action (boy returns to van). The parallel texts given in *Table 4.3* above represent the 'content descriptions' created during the annotation process (see section 4.3 below). The scene contains only minimal dialogue, allowing the AD to 'breathe' and deliver a relatively unhindered audio guide to the action (*Table 4.5*). Although the majority of 'story arcs' selected for inclusion in our corpus contain dialogue in addition to AD, this example illustrates the type of short narrative sequences we sought to isolate. As stated above, our criteria for selecting story arcs (duration, complexity, number of characters present, classes of action etc.) were driven by the current evolutionary state of automated moving image descriptions.

Additional material

In addition to our primary corpus, the AD scripts in the LSMDC data set (Rohrbach et al., 2015), which consists of AD scripts of 180 feature films, will be used to test initial hypotheses about human audio description. Due to its large size, this corpus provides an interesting complementary source of data, although its segmentation is different to the segmentation and extraction of story arcs in our own corpus. The LSMDC data set has been divided into small segments (of approx. 5 seconds in length) and has been annotated with activity elements which were automatically mined from the audio descriptions (Torabi et al. 2016).

As a related undertaking, partners from the University of Helsinki will work closely with YLE archive journalists to gain a comprehensive understanding of the archive content description process. Particular attention will be given to the selection and prioritisation of film materials given over to content description, including factors such as genre specificity and commercial expediency which impact those choices.

4.3 Data Processing and Annotation

Annotation Models and Methods

In parallel with determining the nature of our experimental data, resources were initially focussed on exploring multimodal annotation frameworks. The uncharted nature of future machine description iterations, as the basis of human vs. computer description analyses, required that our annotation methodology was sufficiently flexible to be able to accommodate machine-generated descriptions of varying complexity over the course of the project. Hierarchical multimodal taxonomies (Jimenez & Seibel, 2012) for tagging audiovisual material (narratological, grammatical, and imagery-based), and storytelling ontologies for broadcast news (e.g. BBC (2018) news ontology, *https://www.bbc.co.uk/ontologies/storyline*) were considered as frameworks for annotating semantic and narrative content. However, the former applied tagging protocols that were considerably more numerous and granular than was required for MeMAD purposes (for instance, tagging characters' ages); while the latter, derived from news production workflows, incorporated elements that had no correspondence with feature film analysis (e.g. logging multiple story sources). Hence the early promise of an 'off the shelf' annotation methodology was not realised, and it became apparent that a bespoke methodology would have to be developed.

Based on the theoretical frameworks of discourse processing / storytelling outlined in section 2, we have therefore derived a bespoke annotation model. The starting point in considering the types of annotation that would be required was to conceptualise the highly complex process of multimodal engagement, breaking it down into layers of meaning-making which generally co-occur in the human viewing experience. These are represented in the pyramid featured in Fig. 4.1, whereby in a reading from bottom to top, the level of meaning-making becomes increasingly sophisticated and requires greater cognitive resources in order to retrieve results. Clearly, human understanding transcends a simplistic explanation of the type denoted by a simple 'climbing the ladder' to greater comprehension, but these multiple layers of engagement typify the kinds of human endeavour undertaken in an unspecified and most likely highly individualistic order, in the quest to make sense of complex narrative themes.

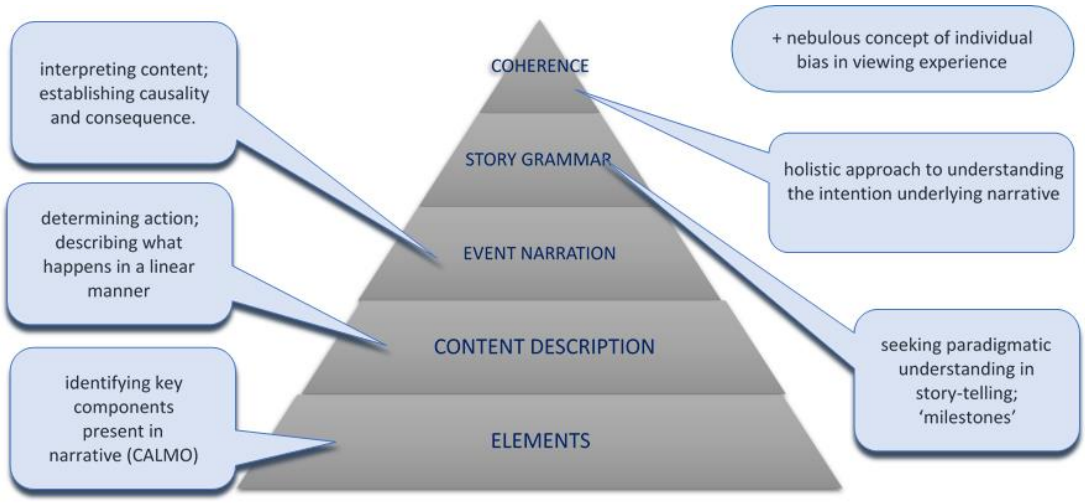


Figure: 4.1 : Accessing multimedia content – Levels of complexity

Hence at the most fundamental level of meaning-making, viewers identify the building blocks of plot exposition, which we identify as **'key elements', (i)**, for the purposes of annotation:

- main characters
- actions
- salient objects
- locations
- the emotional temperature of the piece (mood).

Establishing the nature of these important cues is generally the first task of the viewer, since without a gauge of mood, characterisation and the setting of narrative action, the viewer's inferential skills cannot be fully engaged. Whether or not these initial questions are answered instantly by reference to the film text, the viewer progresses to attempting an understanding of the action taking place, applying other kinds of multimedia cues to facilitate this process. These layers of meaning-making are outlined in the diagram below and matched with their corresponding annotation channels:

(ii) Say what you see (~Content Descriptions): this human activity and corresponding annotation stream represents a 'ground truth' summary of the action taking place on screen; constructed at a descriptive level only (without interpretation), it captures the scene as it would be superficially perceived by the average audience member. In the Relevance Theory (Sperber & Wilson, 1985) model of communication, this corresponds to the level of "what is

said", i.e. the stage before any explicitly or implicitly communicated assumptions have been derived.

(iii) Explicature level (~Event Narration): at this level, a deeper pass of contextual cues is conducted and applied within the wider film context, during which the viewer attempts to establish relevance in relation to particular actions and construct context which in turn informs understanding. In the Relevance Theory model of communication, this corresponds to deriving the *explicature*, i.e. the explicitly communicated assumptions.

(iv) Implicature level (~Story Grammar): may be considered the highest level of narrative immersion, in which key dramatic 'signposts' are assimilated to construct an overarching plot which contains not only points of entry and departure, but also elements of crisis, resolution, failed resolution, and perhaps, conclusion. In the Relevance Theory model of communication, this corresponds to deriving the *implicature(s)*, i.e. the implicitly communicated assumptions.

Our annotations have therefore been designed to address each of these levels of narrative immersion and having been created, will be used as a source of data to evaluate comparable levels of sophistication in machine-generated video descriptions. The flexible nature of the annotations schema means that we are equipped to match any outputs received from the Aalto University computer vision team across the duration of the project. We can also adapt them in order to inform future computer vision models, should machine-generated outputs not match the level of sophistication anticipated at the outset.

Annotation Protocols: Levels 1 and 2

The focus of our 'first pass' annotation process was to create a comprehensive record of the source text data streams: transcripts of the film dialogue were compiled along with their companion audio description scripts. Creating a verbatim dialogue transcript was not originally envisaged, as we expected to use the film screenplays for this purpose. However, screenplays are not universally available, and where they were discovered online, comparisons with the film dialogue suggested they should not be relied upon as a complete and accurate record of the spoken word, many having been derived from fan-sourced material or pre-production drafts later revised during production. Both of the primary data sources (dialogue and AD) represented a departure point for the annotation process although neither, in isolation, can be regarded as a comprehensive resource for the mining of visually salient narrative cues, as outlined above. Furthermore, as a general observation reinforced by the intensive transcription process, although AD may offer a rich seam of visual cues from which certain aspects of narrative might be derived, it should be kept in

mind that on the textual surface AD is no more than a partial representation of the audiovisual content to which it refers (section 2.3). The textual surface of AD serves as a starting point for creating a comprehensive mental representation of the audiovisual content, but it cannot be regarded as the sole source of narrative saliency. For this reason, we rejected the idea of a direct comparison between audio description and machine-generated video descriptions, which we determined to be methodologically flawed.

Key Elements

Having established that the combined data from AD and dialogue streams would not suffice for the purposes of comparing human- and machine-generated video descriptions, we explored supplementary annotation protocols. Since existing data streams could not deliver a comprehensive description of narrative, we elected to create our own. The first step in this process was to compile a list of narrative building blocks found in all film material, i.e. rudimentary components of plot which would be identified as relevant to meaning-making by the average viewer. We termed these components 'key (dramatic) elements', as they comprised: ***character*** (e.g. man, woman, young girl, small boy), ***action*** (e.g. sitting, walking, talking, eating), ***location*** (e.g. at the office, in the kitchen, on a road), ***mood*** (happy, sad, angry etc.), ***action-relevant object*** (e.g. car, desk, bed) and optionally, ***gestural/body language*** (a shrug, a pointing finger). The value of extracting 'key elements' as an entry point to the annotation and analysis process is that they are the *sine qua non* of all dramatic texts. Although all of these elements may not be present at any given juncture, a combination of two or more will generally be critical to plot development and exposition and can therefore be regarded as narratively important.

Content Description

Moving beyond simple identification of key elements and acknowledging the need for a rudimentary description of film action which expands on the partial descriptions provided by the AD, we adopted a *ground truth* annotation, which we termed the 'content description' (CD). The purpose of employing this secondary stratum of annotation was to establish a factual description of the action occurring on screen while avoiding incursions into interpretation, in order to safeguard objectivity. Issues of causality and consequence in relation to narrative actions were therefore excluded as far as possible from 'content descriptions', these aspects being reserved for higher level annotations (below).

A sample content description taken from our annotation of *Little Miss Sunshine*, reads: "Olive and Dwayne stand up and slowly walk towards the bottom of the slope" (section 4.1). Evidently, as suggested by this annotation, content descriptions are based on a 'say what you see' strategy, offering a means of extracting elements which a human viewer would

recognize as story-sensitive, while affording those elements minimal narrative context. As such, this *ground truth* is not available from any other source, and represents a 'plain vanilla', factual representation of events. The intention is to compare this literal text with the equally literal (non-interpretive) computer-generated video descriptions from early machine iterations which tend to be similarly descriptive rather than interpretive.

Annotation protocols: Levels 3 and 4

Mental modelling frameworks and theories of relevance in meaning-making (section 2.1) suggest that we interpret patterns of speech and observed behaviours by identifying pertinent cues from a barrage of visual and audio cues found in multimedia materials, arranging these in multiple possible permutations (mental models) until we arrive at an explanation that is the most natural and plausible (optimally relevant) according to our best abilities. Moving on from basic comprehension of events to interpretation and conjecture requires the viewer to employ 'extradiegetic' references such as social convention, cultural norms and life experience. Matching the output of this task requires a different approach to annotation, involving interpretation and narrative mapping. These elements are mirrored in two further levels of annotation which we have termed 'event narration' and 'story grammar' (Mandler & Johnson, 1977; Mandler, 1984; Appose & Karrupali, 1980).

Event Narration

As noted above, the 'event narration' (EN) annotation stream broadly corresponds to the explicature level in Relevance Theory. In this respect, event narration extends beyond the surface text of content description, seeking to address issues of causality and (local) consequence. EN seeks to contextualise events within the micro-narrative at the centre of the story arc, cross-referencing possible inferences from outside the story arc, and yet not, at this stage, attempting to construct an 'aerial view' of the entire plot. Effectively, the EN annotations record the 'why' for events occurring in the narrative, and explicate cohesive links across the wider storyline.

Story Grammar

Both the fact-based, say-what-you-see 'content descriptions' and the 'event narrations' (an interpretive stream of annotation already incorporated into our movie dataset, see above) allow us to determine which elements of audiovisual narrative contribute to coherent storytelling and plot exposition within each of our previously isolated story arcs. These annotation streams are available to supplement the 'character-action-location-mood-object' tags entertained at the more fundamental level of analysis of human vs. machine comparison (see section 4).

However, if machine-based audiovisual coherence descriptors prove sufficiently robust, and there is evidence of computer-generated story arc exposition, we envisage re-visiting our human annotated corpus and selecting a representative sample of video extracts in order to apply 'story grammar' tagging (Mandler & Johnson, 1977; Mandler, 1978). These annotations would be appended to critical intersections in the exposition of narrative, flagging up key milestones such as initiating event, internal response, plan, attempt to enact plan, consequence and reaction (Appose & Karuppali, 1980:4; see also section 2.1). Referencing theoretical frameworks and the impact of Relevance Theory (Sperber & Wilson, 1985), this path to story resolution produces an 'implicature' that is readily derived from a summary of audio and visual cues, seen through the eyes of a sentient being endowed with pragmatic world knowledge.

In the event that automated audiovisual cue extraction fails to produce narratively coherent machine descriptions at a macrostructural level during the life of the project, 'story grammar' annotations can be analysed from within the human-generated film corpus, as a means of determining the manner in which human understanding of plot extends beyond that of the most advanced computer vision models.

A summary of each annotation category is shown in *Fig. 4.1* below.

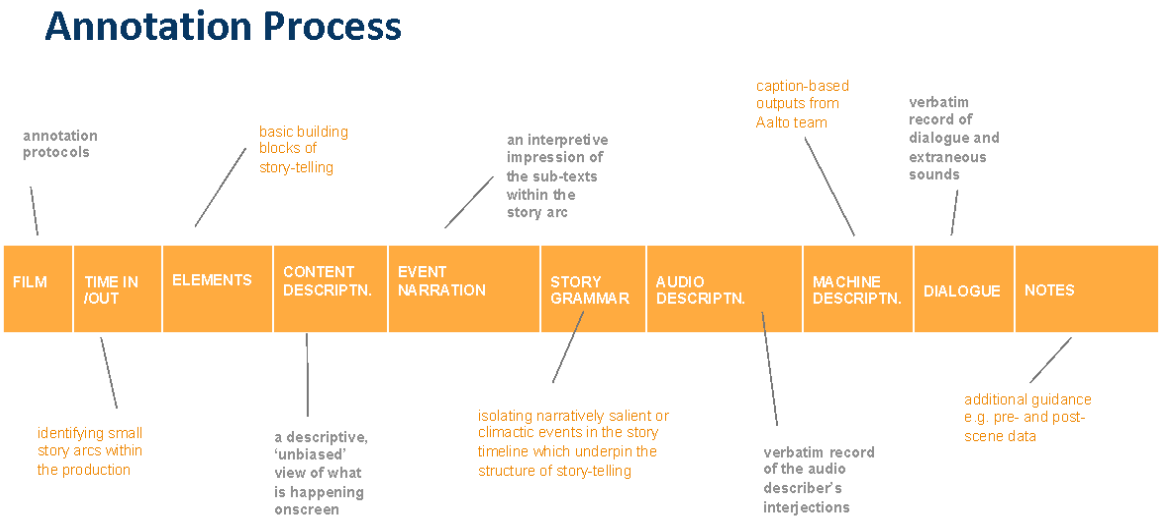


Figure 4.2: Annotation Categories

Annotation Workflow

The initial annotation process was undertaken by doctoral and post-doctoral researchers at the University of Surrey who are experienced in multimodal analysis and/or audio description. Annotators began by viewing each film in its entirety, in order to gain an

appreciation of the broad narrative structure of the piece. This initial viewing was combined with 'spotting' for story arcs (noting time-in and time-out) which met the criteria described above. In order that future machine descriptions could be fairly compared with their human annotation counterparts, these short extracts were selected to stand alone in terms of narrative completeness. However, it is acknowledged that access to the wider narrative significance of these brief 'story-arcs' may be found in cues which lie outside the extract, occurring either earlier, or indeed later, in the exposition of the film. Attempts to mitigate any insights lost to this effect were addressed in the construction of 'event narration' annotations, where the interpretation of micro-plots by reference to wider narrative strategies was captured (see above). In 'spotting' mode, our annotators simply identified suitable story arcs, continuing to watch the film in a linear fashion throughout this process. This ensured that the holistic viewing experience was not compromised by a need to pause and complete annotations after each 'story arc' had been selected. Having completed this task, our annotators returned to the first of the selected extracts and began the annotation activity. At this point, extracts were revisited in order of occurrence in the film presentation, capturing dialogue, AD script, 'key elements', 'content descriptions' and 'event narration' in one pass.

Validity of Human Annotations

Human beings make sense of the world from their own unique perspective. We apply individual life experience, personal prejudice and bias, lessons adapted from formal education, an innate and personal moral compass, the results of earlier 'trial and error' approaches in problem-solving, and intuition to navigate the innumerable cues that require decoding for the purposes of meaning-making. Naturally, this highly individualistic perspective can prove problematic where human operatives are required to perform a qualitative task in a standardized and uniform manner. Accepting that absolute standardization in these circumstances is realistically beyond reach, we established a set parameters to minimise variation in our human-generated annotations. These guidelines captured the description of 'mood', the treatment of 'location' and the selection of narratively salient 'objects', for instance.

Levels of granularity in description-writing also call for a uniform approach, with the example of whether one sees, for example, an animal, a dog or a Scottish terrier as being pertinent both to the human annotation schema, and in setting expectations for our comparisons with the machine descriptions. Hence, hypernyms, hyponyms and synonyms will be considered in terms of their inter-relatedness within the Wordnet (synset) concept. Future work exploring acceptable tolerance levels across related words will be required to resolve this issue.

At the time of writing, user-testing of our annotation system has been initiated with archive journalists at YLE broadcasters in Finland. Following meetings with key staff in early December, a short series of film extracts and accompanying evaluative questionnaire have been distributed amongst the video content descriptions team. We expect to receive their responses early in January 2019. Finally, it is our intention to undertake inter-rater reliability evaluations on human annotation outputs, assuming there is sufficient available resource across the duration of the project. However, it is not certain at this stage whether this might prove too labour-intensive an undertaking.

4.4 Example of Annotation

In 'Elements' yellow highlights those aspects present in the MD.
In 'Event Narration' pink indicates human interpretation contextualising dialogue and actions.

Frame/Time codes	Machine Description (MD)	Audio Description (AD)	Elements (E)	Content Description (CD)	Event Narration (EN)
02:100994/01:07:19.760 	a man is sitting in a field		C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.	Dwayne is sitting on the grass in a field, hugging his knees. He is sitting with his back to us.	Dwayne is upset.

02:101125/01:07:25 	a man and a woman are talking to each other	He is sitting with his back to her, arms resting on his knees, gazing at the rocky soil at his feet, and doesn't turn as she comes near.	C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.	Olive walks towards Dwayne, who is sitting on the ground, staring at the grass. Sheryl, Frank and Richard are at the top of the slope, standing next to the van, looking down at them.	Dwayne is very upset: his dreams have been shattered … he just discovered that he is colour-blind and cannot fly fighter jets.
02:101650/01:07:46.000 	a group of people are singing and dancing	Dressed in her red T-shirt, pink shorts and red cowboy boots, her long hair tied back, her huge glasses perched on her nose, Olive squats at Dwayne's side.	C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.	Once she has reached Dwayne, Olive slows down and bends her knees to sit next to Dwayne. Dwayne does not react.	Olive is sad for her brother and wishes to reassure him. She looks slightly worried at how he might react to her presence and touch.

02:101875/01:07:55.0  	a group of people are in a field		C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.		
02:102325/01:08:13.000 	a man is running	Olive stands up and Dwayne gets to his feet and goes with her to the bottom of the slope.	C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.	Olive and Dwayne stand up and slowly walk towards the bottom of the slope.	

02:102475/01:08:19.000 	a man and a woman are walking in a field		C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.		
02:102625/01:08:25.000 	a woman is walking down the road		C: A boy; a little girl. A: Sitting, walking, hugging, climbing. L: Field (road) O: Field, grass M: Sad Oth: (Gesture) Hug.		

Table 4.4: Example of Annotation - Little Miss Sunshine ('Boy in a field')

4.5 Planned analytical steps

In addition to creating the audiovisual corpus and the annotations as described above, we have also explored different ways of analysing the data. Mirroring the multi-layered approach to creating annotations for the film corpus extracts, our analysis will take a similarly stratified path. Drawing on the theoretical frameworks of human meaning-making (section 2), the analytical process is designed with inherent agility in order to handle expected increments in the convolution of computer-generated descriptions. It also reflects the complex strategies for plot assimilation adopted by human audiences of film narrative. This is illustrated in the 'knowledge pyramid' shown in *Fig. 4.1* and consists of a number of layers of understanding upon which individuals draw when attempting to access story narrative in multimodal material. The layers correspond to the protocols in our annotation schema.

Iterative Processing

In order to make meaningful comparisons between the human and machine-generated descriptions, an iterative approach will be adopted.

Before initiating a comparative analysis of human- and machine-generated data it will be necessary to determine which of the currently available computer vision training datasets is most appropriate for producing a first iteration (best efforts) of our machine-generated descriptions. The Aalto University team will test, for accuracy, a representative sample of currently available datasets early in 2019, systematically applying test data and standard evaluation metrics (e.g. METEOR, CIDEr, BLEU) to measure the efficacy of outputs. Initial evaluations are to be conducted with minimal human intervention, applying superficial statistical testing of the quality of descriptions, until the final stages of dataset refinement, when a more in-depth analysis involving human inspection of the data sets will be required. Time permitting, Surrey could also provide a comparative analysis of shortlisted datasets from the human annotation perspective.

Iteration 1

Once we have established the dataset(s) that is (are) most fit for purpose, first iteration machine- generated descriptions will be produced from our annotated corpus of fifty feature films.

Comparative lexical studies will be an important analytical step for the first iteration of machine descriptions. We aim to make a comparative lexical study of audio descriptions and these machine-generated descriptions, seeking out differences in patterns of word use, informativeness values, omissions and misrepresentations. As moving image descriptions focus on the actions at the heart of each narrative, our intention is to concentrate, initially,

on verbs and verbal phrases, drawing out evidence of differences in approach and outputs between corpora.

Frequency analysis, including high and low occurrences of particular lexical terms (unique words and phrases, those which are over-frequently used in relation to lexical norms etc.) will be used to identify areas of interest, and examples subsequently selected for qualitative analysis on a case-by-case basis. As pointed out above, the LSMDC16 corpus of audio descriptions will be consulted for reference purposes to test hypotheses arising from our own data sets.

Regarding the qualitative analyses, we expect these to involve some deeper understanding of the material comprising the machine-generated training data from which the computer outputs are drawn, since this may inform certain expected anomalies within our results.

Presence of key visual elements. Because the pilot stage showed that a direct comparison between AD and MD makes little sense, our basic annotation layers, i.e. the key elements and our content descriptions, will be used as a 'ground truth' for the comparison and evaluation of the two parallel corpora. Rather than expecting either of these corpora to provide complete descriptions of the visual images in the video clips, we will explore whether the key elements that we identified in the video clips are represented in each of the two the sub-corpora and to what extent the lexical choices that were made for describing the key elements are accurate representations of the elements.

Subsequent Iterations

Contingent upon the results of the first iteration, it is currently projected that a further two machine-generated computer description iterations could be delivered by the Aalto University computer vision team during the period M18-24. However, this scenario remains fluid with the results of the first iteration still several months away, and a final decision will be taken about the shape and number of future iterations at M18. In particular, since the first machine iteration is regarded as the 'current state of the art' as far as computer-produced video descriptions are concerned, it should logically serve as a benchmark for measuring future progress in the development of machine descriptions across both tasks 5.1 and 5.2. Given the level of resource available and the time-intensive nature of human vs. machine descriptions analysis, we anticipate analysing additional iterations, in excess of those projected below, would not be feasible during the life of the project.

Notionally, future iterations might comprise:

(i) incrementally enhanced machine descriptions which draw on Iteration 1. above and, additionally, introduce sequence modelling techniques to mimic visual coherence between film frames, drawing on the work outlined in the VIST (Huang et al., 2016) and LSMDC (Rohrbach et al., 2015) studies (Iteration 2); and

(ii) incrementally enhanced machine descriptions which draw on Iteration 1. and 2. above, to include sequence modelling, with the addition of audio segmentation and diarisation techniques (see section 3.3), i.e. extraction of sound features to measure impact, if any, on increasing inter-frame coherence (Iteration 3).

Assuming this machine-driven iterative programme were to be delivered, specifically in relation to sequence modelling, with or without the inclusion of audio information, an increasingly complex association of ideas between frames presented in corresponding machine description outputs would allow for a more sophisticated level of analysis and interpretive comparison to be undertaken with human annotations. We anticipate that a smaller sample of human-generated annotations would be re-visited in this case, and story grammar 'milestones' (Appose & Karrupali, 1980) added to our original annotations schemata (section 3.2 above), to denote key moments of narrative storytelling and action-based inter-relatedness between contiguous image frames. This would enable a comparison between machine sequence-modelled story arcs and their human-annotated parallel texts, with particular attention being paid to instances of co-occurrence or omission. Narratively intentional words and phrases in the machine-derived lexicon ('next', 'because', 'then', 'due to' etc.) and repetition of key iconographical indicators (e.g. 'meeting', 'birthday', 'holiday', 'graduation') should point to evidence of a predetermined story 'macrostructure' (Appose & Karuppali, 1980:1). These concepts elide with Mandler's notion of cognitive schemata (see section 2.1), upon which the comprehension of narrative is contingent, and which subsume storyline expectations, plot units, the sequencing of narrative and the interconnectivity between story components.

Hence, the agility of the annotation system we have adopted lends itself to adaptation for any complexity-level of machine outputs envisaged during the life of the project. However, in the event that the level of sophistication achieved by the machine descriptions fails to deliver internally coherent storytelling, an investigation of computer shortcomings would be used to inform future iterations, assessing key differences between human and machine recognition of intertextual referencing via the 'milestones' approach cited above. Detailed plans for this work will be explored between M12 and M18, and developed further.

Appendix I: Dissemination Activities and User Testing

Dissemination Activities

Conference presentation: Languages and the Media, Berlin, 3-5th October 2018	Sabine Braun and Kim Starr presented a paper entitled: "*From Slicing Bananas to Pluto the Dog: Human and Automatic Approaches to Visual Storytelling.*"
Blogpost and Twitter feed dissemination	'From Slicing Bananas to Pluto the Dog: Computer Vision with a Human Touch', written by Kim Starr **https://memad.eu/2018/10/19/slicing-bananas-pluto-dog-computer-vision-human-touch/**
Conference presentation: Presentation to TECHNE's (UK Research Councils Group) Annual Congress June 2018	Kim Starr presented her post-doctoral work on the MeMAD project to TECHNE (UKRI-funded) scholars, academics and representatives from supporting industrial partners, at the annual research convention.
Conference presentation: ARSAD 2019, Barcelona, March 2019	Paper by Sabine Braun and Kim Starr accepted at audio description specific international conference attended by academics, broadcast industry professionals and parties with a particular interest in visual accessibility.
Conference presentation: Media4All, Stockholm, June 2019	Paper by Sabine Braun and Kim Starr accepted at international media accessibility conference, covering audio description, subtitling, dubbing, signing and all aspects of media accessibility. Attended by similar audience to ARSAD, but with broader accessibility remit.
Book publication *'Innovations in Audio Description'*, 2019	Volume commissioned by academic publishers Taylor & Francis/Routledge, co-edited by Sabine Braun and Kim Starr. To include chapter on MeMAD human vs. machine annotation and analysis methodology (authors: Braun, Starr, Hirvonen, Laaksonen, Tiittula).

User Testing

Sample film extracts and semi-structured questionnaire used to garner feedback on accuracy and completeness of human-generated content annotations from television industry professionals.	Audience: YLE archive journalists (ongoing)

References

AENOR Standard UNE 153020 (2005) *Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías*. Madrid: AENOR.

Appose, A. and Karuppali, S. (1980) 'Decoding the Macrostructural Form of Oral Narratives in Typically Developing Children Between 6 - 11 Years of Age: Using Story Grammar Analysis'. *Online Journal of Health and Allied Services,* 17 (1), article 12. Available at: https://www.scopus.com/record/display.uri?eid=2-s2.0-85047524895&origin=inward&txGid=979609e35b955680098849bcea1fd82a. Accessed on 17th December, 2018.

Arnold, D. and Whitney, D. (2005) 'Adaptation and Perceptual Binding in Sight and Sound', in *Fitting the Mind to the World,* by Clifford, C. and Rhodes, G. (ed.) pp. 339-360. Oxford: OUP.

Babaee, M., Dinh, D.T. and Rigoll, G. (2018) 'A Deep Convolutional Neural Network for Video Sequence Background Subtraction'. *Pattern Recognition*, 76, pp. 635-649.

Bal, M. (2009) *Narratology: Introduction to the Theory of Narrative.* 3rd edn. Toronto: University of Toronto Press.

Bal, M. and Lewin, J. (1983) 'The Narrating and the Focalizing: A Theory of Agents in Narrative'. *Style*, 17 (2), pp. 234-269.

Bardini, F. (2017) 'Audio Description Style and the Film Experience of Blind Spectators: Design of a Reception Study'. *Rivista Internazionale di Tecnica della Traduzione / International Journal of Translation* (19), pp. 49-73.

Bartlett, F.C. (1932) *Remembering: A Study in Experimental and Social Psychology.* New York, NY, USA: Cambridge University Press.

Baumgarten, N. (2008) 'Yeah, that's it!: Verbal reference to visual information in film texts and film translations'. *Meta*. 53 (1), pp. 6-25.

Bell, V. and Johnson-Laird, P. (1998) 'A Model Theory of Modal Reasoning.' *Cognitive Science*, 22, pp. 25-51.

Benecke, Bernd (2014) *Audiodeskription als partielle Translation - Modell und Methode*. Berlin: LitVerlag.

Blakemore, D. (1992) *Understanding Utterances.* Oxford: Blackwell.

Bourne, J. and Jiménez, C. (2007) 'From the visual to the verbal in two languages: a contrastive analysis of the audio description of *The Hours* in English and Spanish', in Dìaz-Cintas, J., Orero, P. and Remael, A. (eds.) *Media for All: Subtitling for the Deaf, Audio Description and Sign Language.* Amsterdam: Rodopi, pp. 175[SKD(&L1] -187.

Branigan, E. (1984). *Point of view in the cinema*. Berlin: De Gruyer.

Braun, S. (2007) 'Audio Description from a Discourse Perspective: A Socially Relevant Framework for Research and Training.' *Linguistica Antverpiensia*, NS6. pp. 357-369.

Braun, S. (2011) 'Creating Coherence in Audio Description'. *Meta*, 56 (3), pp. 645-662.

Braun, S. (2016) 'The Importance of Being Relevant? A cognitive-pragmatic framework for conceptualising audiovisual translation'. *Target*, 28 (2), pp. 302-313.

Bogucki, L. (2004) 'The Constraint of Relevance in Subtitling'. *JosTrans,* (1), pp. 71-88.

British Broadcasting Corporation (2018) *Storyline Ontology*. Available online at: https://www.bbc.co.uk/ontologies/storyline. Accessed on 19th December, 2018.

Brown, G. and Yule, G. (1983) *Discourse Analysis*. Cambridge: CUP.

Bublitz, W. (1999) 'Introduction: Views of Coherence', in Bublitz, W., Lenk, U. and Ventola, E. (eds.) *Coherence in Spoken and Written Discourse*. Amsterdam and Philadelphia: John Benjamins, pp. 1-7.

Bublitz, W. and Lenk, U. (1999) 'Disturbed coherence: 'Fill me in'', in Bublitz, W. Lenk, U. and Ventola, E. (eds.) *Coherence in Spoken and Written Discourse*. Amsterdam and Philadelphia: John Benjamins, pp. 153-174.

Bublitz, W., Lenk, U. and Ventola, E. (1999) *Coherence in Spoken and Written Discourse: How to Create it and How to Describe it*. Amsterdam and Philadelphia: John Benjamins.

Chaume, F. (2004) *Cine y Traducción*. Madrid: Cátedra.

Chen, D. and Dolan, W. (2011) 'Collecting highly parallel data for paraphrase evaluation'. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* 1, pp. 190-200.

Council Directive 2010/13/EC of the European Parliament and of the Council of 10 March 2010 on the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States Concerning the Provision of Audiovisual Media Services (Audiovisual Media Services) (2010). Official Journal of the European Communities, L 95/1-24. Available at: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0013. Accessed on 20th December, 2018.

Davila-Montess, J. and Orero, P. (2016) 'Audio Description Washes Brighter? A Study in Brand Names and Advertising', in Matamala, A. and Orero, P. (eds.) *Researching Audio Description, New Approaches*. London: Palgrave Macmillan, pp. 123-142.

De Beaugrande, R. and Dressler, W. (1981) *Introduction to Text Linguistics.* London: Longman.

Desilla, L. (2012) 'Implicatures in Film: Construal and Functions in Bridget Jones romantic comedies.' *Journal of Pragmatics* 44 (1), pp. 30-53.

Diaz Cintas, J. and Remael, A. (2007) *Subtitling*. Manchester: St. Jerome.

Dicerto, S. (2018) *Multimodal Pragmatics and Translation: A New Model for Source Text Analysis.* London: Palgrave Macmillan.

Fels, D., Udo, J-P., Diamond, J. and Diamond, J. (2006) 'Comparison of Alternative Narrative Approaches to Video Description for Animated Comedy.' *Journal of Visual Impairment and Blindness* 100 (5), pp. 295-305.

Forceville, C. (2014) 'Relevance Theory as a model for multimodal communication', in Machin, D. (ed.) *Visual Communication*. Berlin: De Gruyter Mouton, pp. 51-70.

Fresno, N. (2014) *La (re)construcción de los personajes fílmicos en la audiodescripción.* PhD thesis. Universitat Autònoma de Barcelona. Available at: http://www.tdx.cat/bitstream/handle/10803/285420/nfc1de1.pdf. Accessed on 17th December, 2018.

Fresno, N. (2016) 'Carving characters in the mind. A theoretical approach to the reception of characters in audio described films'. *Hermeneus, TI*, 18, pp. 59-92.

Frida (2002) Directed by Julie Taymor [Film]. UK: Lions Gate Home Entertainment, UK.

Fryer, L. (2016) *An Introduction to Audio Description, A Practical Guide.* Abingdon, Oxon: Routledge.

Gernsbacher, M. and Talmy, G. (1995) *Coherence in Spontaneous Text.* Amsterdam: Benjamins.

Gutt, E-A. (2000) *Translation and Relevance: Cognition and Context.* Manchester: St Jerome Publishing.

Halliday, M. A. K. (1985) *An Introduction to Functional Grammar.* London: Edward Arnold.

Halliday, M. A. K. and Hasan, R. (1976) *Cohesion in English*. London: Longman.

He, K., Zhang, X., Ren, S., and Sun, J. (2016) 'Deep Residual Learning for Image Recognition'. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*. Available at: https://arxiv.org/abs/1512.03385. Accessed on 14th December, 2018.

Herman, D. (2002) *Story Logic*. Lincoln: University of Nebraska Press.

Herman, D. (2013) *Cognitive Narratology*. Available online at: http://www.lhn.uni-hamburg.de/article/cognitive-narratology-revised-version-uploaded-22-september-2013. Accessed on 19th December, 2018.

Hirvonen, M. (2012) 'Contrasting Visual and Verbal Cueing of Space: strategies and devices in the audio description of film'. *New Voices in Translation Studies*, 8, pp. 21-43.

Hirvonen, M (2013a) 'Perspektivierungsstrategien und -mittel kontrastiv: Die Verbalisierung der Figurenperspektive in der deutschen und finnischen Audiodeskription'. *trans-kom: Zeitschrift für Translationswissenschaft und Fachkommunikation,* 6 (1), pp. 8-38.

Hirvonen, M. (2013b) 'Sampling Similarity in Image and Language – Figure and Ground in the Analysis of Filmic Audio Description'. *SKY Journal of Linguistics,* 26, pp. 87-115.

Hirvonen, M. and Tiittula, L. (2012) 'Verfahren der Hörbarmachung von Raum. Analyse einer Hörfilmsequenz', in Hausendorf, H., Mondada, L. and Schmitt, R. (eds.) *Raum als interaktive Ressource*. Tübingen: Narr, pp. 381-427.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short Term Memory'. *Neural Computation*, 9 (8), pp. 1735-1780.

Huang, T. H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Dhruv, B., Zitnick, C., Parikh, D., Vanderwende, L., Galley, M. and Mitchell, M. (2016) 'Visual Storytelling'. *Proceedings of NAACL-HLT,* San Diego, California, June 12-17, pp. 1233-1239.

Hyks, V. (2005) 'Audio Description and Translation, Two Related but Different Skills'. *Translating Today*, 4, pp. 6 - 8.

Ibanez, A. (2010) 'Evaluation Criteria and Film Narrative. A Frame to Teaching Relevance in Audio Description'. *Perspectives: Studies in Translatology,* 18 (3), pp. 143-153.

Independent Television Commission (2000) Guidance on Standards for Audio Description. Available at:audiodescription.co.uk/uploads/general/ itcguide_sds_audio_desc_word3.pdf. Accessed on 18th December, 2018.

Jamieson, H. (2007) *Visual Communication: More Than Meets the Eye*. Bristol: Intellect.

Johnson-Laird, P. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge/Mass.: Harvard University Press

Johnson-Laird, P. (2006) *How We Reason*. Oxford: OUP.

Kim, T., Heo, M-O., Son, S., Park, K-W., Zhang, B-T. (2018) *GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation* Available online at: https://arxiv.org/abs/1805.10973. Accessed on 18th December, 2018.

Kovačič, I. (1993) 'Relevance as a Factor in Subtitling Reduction', in Dollerup, C. and Lindegaard, A. (eds.) *Teaching Translation and Interpretation 2: Insights, Aims, Visions*. Amsterdam: Benjamins, pp. 245-251.

Kress, G. (1998) 'Visual and Verbal Modes of Representation in Electronically Mediated Communication', in Snyder, I. and Joyce, M. (eds.) *Page to Screen*. Sydney: Allen & Unwin, pp. 53-79.

Krishna, R., Zhu, Y.,Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L-J., Shamma, D., Bernstein, M.S., Li, F-F., (2017) 'Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations'. *International Journal of Computer Vision*, 123, pp. 32-73.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012) 'Imagenet classification with deep convolutional neural networks'. *Advances in Neural Information Processing Systems,* pp. 1097–1105.

Kruger, J. L. (2010) 'Audio Narration: Re-Narrativising Film'. *Perspectives: Studies in Translatology*, 18, pp. 231-249.

Lemke, J. (2006) 'Toward critical Multimedia Literacy: Technology, Research, and Politics' in McKenna, M. (ed.) *International Handbook of Literacy and Technology*, 2. Mahwah/NJ: Erlbaum, pp. 3-14.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollar, P. (2015) 'Microsoft COCO: Common Objects in Context'. *Computer Vision,* ECCV 2014, pp. 740–755.

Little Miss Sunshine (2006) Directed by Jonathon Dayton and Valerie Faris [Film]. USA: 20th Century Fox Home Entertainment.

Lopez, M., Kerney, G. and Hofstädter, K. (2018) 'Audio Description in the UK: What works, What doesn't, and Understanding the Need for Personalising Access'. *British Journal of Visual Impairment*, 36 (3), pp. pp. 274–291.

Mälzer-Semlinger, N. (2012) 'Narration or description: What should Audio Description "look" like?', in Perego, E. (ed.) *Emerging topics in translation: Audio Description*. Trieste: Edizioni Università di Trieste, pp. 29-36.

Mandler, J. (1978) 'A Code in the Node'. *Discourse Processes*, 1 (1), pp. 14-35.

Mandler, J. (1984) *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Hillsdale, NJ: Lawrence Erlbaum.

Mandler, J. and Johnson, N. (1977) 'Remembrance of Things Parsed: Story Structure and Recall'. *Cognitive Psychology*, 9, pp. 111-151.

Mangiron, C. and Zhang, X. (2016) 'Game Accessibility for the Blind: Current Overview and the Potential Application of Audio Description as the Way Forward', in Matamala, A. and Orero, P. (eds.) *Researching Audio Description*. London: Palgrave Macmillan, pp. 75-95.

Martínez Sierra, J. J. (2010) 'Approaching the Audio Description of Humour.' *Entreculturas: revista de traducción y comunicación intercultural*, 2, pp. 87-103.

Matamala, A. (2018) 'One Short Film, Different Audio Descriptions. Analysing the Language of Audio Descriptions Created by Students and Professionals. *Onomazein*, 41, pp. 186-207.

Mazur, I. and Chmiel, A. (2012) 'Audio Description Made to Measure: Reflections on Interpretation in AD Based on the Pear Tree Project Data', in Remael, A., Orero, P. and Carroll, M. (eds.) *Audiovisual Translation and Media Accessibility at the Crossroads: Media for All 3.* Amsterdam: Rodopi, pp. 173-188.

Mubenga, K.S. (2009) 'Towards a Multimodal Pragmatic Analysis of Film Discourse in Audiovisual Translation'. *Translators' Journal* (54: 3), pp. 466-484.

Myers, J. L., Cook, A., Kambe, G., Mason, R. and O'Brien, E. (2010) 'Semantic and Episodic Effects on Bridging Inferences'. *Discourse Processes*, 29 (3), pp. 179-199.

O'Halloran, K., Tan, S. and Marissa, K. (2014) 'Multimodal Pragmatics', in Schneider, K. and Barron, A. (eds.) *Pragmatics of Discourse*. Berlin: De Gruyter Mouton, pp. 239-268.

Orero, P. (2011) 'Audio Description for Children: Once upon a time there was a different audio description for characters'. In Di Giovanni, E. (ed.) *Entre texto y receptor: accesibilidad, doblaje y traducción.* Frankfurt: Peter Lang, pp. 169-184.

Ramos Caro, M. (2016) 'Testing Audio Narration: the Emotional Impact of Language in Audio Description'. *Perspectives: Studies in Translation Theory and Practice,* 24 (4), pp. 606-634.

Ren, Z., Wang, X., Zhang, N., Lv, X. and Li, L-J. (2017) 'Deep Reinforcement Learning-based Image Captioning with Embedding Reward'. Available online at: https://arxiv.org/abs/1704.03899. Accessed on 18th December, 2018.

Rohrbach, A., Rohrbach, M., Tandon, N. and Schiele, B. (2015) 'A dataset for movie description'. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Available at: http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Rohrbach_A_Dataset_for_2015_CVPR_paper.pdf. Accessed on 14th December, 2018.

Rohrbach, A., Rohrbach, M., Tang, S., Oh, S. J. and Schiele, B. (2017) 'Generating descriptions with grounded and co-referenced people'. *Proceedings of the* IEEE *Conference on Computer Vision and Pattern Recognition*. Available online at: https://arxiv.org/abs/1704.01518. Accessed on 20th December, 2018.

Salway, A. (2007) 'A Corpus-based analysis of the language of audio description', in Dìaz Cintas, Orero, P. and Remael, A. (eds.) *Media for all : Subtitling for the Deaf, Audio Description and Sign Language.* Amsterdam and New York: Rodopi, pp. 151-174.

Shank, R. C., and Abelson, R. (1977). *Plans, scripts, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Sharma, P., Ding, N., Goodman, S. and Soricut, R. (2018) 'Conceptual Captions: A Cleaned, Hypernymed, Image alt-text Dataset for Automatic Image Captioning'. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL 2018, Vol. 1, Melbourne, Australia, July 15-20th, pp. 2556–2565.

Smilevski, M., Lalkovski, I. and Madjarov, G. (2018) 'Stories for Images-in-Sequence by using Visual and Narrative Components'. *Communications in Computer and Information Science*, 940, pp. 148-159.

Sperber, D. and Wilson, D. (1995) *Relevance: Communication and Cognition*. 2nd edn. Oxford: Blackwell.

Starr, K. (2018) *Audio Description and Cognitive Diversity: a bespoke approach to facilitating access to the emotional content in multimodal narrative texts for autistic audiences*. PhD thesis. University of Surrey. Available at: http://epubs.surrey.ac.uk/848660/.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D. Vanhoucke, V. and Rabinovich, A. (2015) 'Going Deeper with Convolutions'. *Proceedings of the IEEE conference 2015 on Computer Vision and Pattern Recognition.* Available online at: https://arxiv.org/abs/1409.4842. Accessed on 20th December, 2018.

Talmy, L. (1983) *How Language Structures Space.* New York: Plenum Press.

Tanskanen, S-K. (2006) *Collaborating towards Coherence: Lexical Cohesion in English Discourse*. Amsterdam: Benjamins.

The Hours (2003) Directed by Stephen Daldry [Film]. USA: Disney.

Torabi, A., Tandon, N., and Sigal, L. (2016) *Learning Language – Visual Embedding for Movie Understanding with Natural Language.* Available online at: https://arxiv.org/abs/1609.08124. Accessed on 19th December, 2018.

Vallet, F., Essid, S. and Carrive, J. (2013) 'A Multimodal Approach to Speaker Diarization on TV Talk-Shows'. *IEEE Transactions on Multimedia*, 15 (3), pp. 503-520.

Vandaele, J. (2012) 'What Meets the Eye. Cognitive Narratology for Audio Description'. *Perspectives: Studies in Translatology,* 20 (1), pp. 87-102.

van Dijk, T. and Kintsch, W. (1983) *Strategies of Discourse Comprehension*. New York: Academic Press.

Venugopalan, S., Rohrback, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. (2015) 'Sequence to Sequence - Video to Text'. *Proceedings of 2015 IEEE International Conference on Computer Vision*. Available online at: https://arxiv.org/abs/1505.00487. Accessed on 18th December, 2018.

Vercauteren, G. (2007) 'Towards a European Guideline for Audio Description', in Dìaz-Cintas, J., Orero, P. and Remael, A. (eds.) *Media for All: Subtitling for the Deaf, Audio Description, and Sign Language.* Amsterdam: Rodopi, pp. 139-150.

Vercauteren, G. and Remael, A. (2014) 'Audio-describing Spatio-Temporal Settings', in Orero, P., Matamala, A. and Maszerowska, A. (eds.) *Audio description: New Perspectives Illustrated*. Amsterdam: Benjamins, pp. 61-80.

Walczak, A. and Fryer, L.(2017) 'Creative Description: The Impact of Audio Description on Presence in Visually Impaired Audiences'. *British Journal of Visual Impairment,* 35 (1), pp. 6-17.

Wilken N. and Kruger, J. (2016) 'Putting the Audience in the Picture: Mise-en-Shot and Psychological Immersion in Audio Described Film'. *Across Languages and Cultures,* 17 (2), pp. 251-270.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Gao, Q., Macherey, K. (2016) *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Available online at: https://arxiv.org/abs/1609.08144v2. Accessed on 18th December, 2018.

Xu , J., Mei, T., Yao, T. and Rui, Y. (2016) 'MSR-VTT: A Large Video Description Dataset for Bridging Video and Language'. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288-5296.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C. Larochelle, H. and Courville, A. (2015) *Describing Videos by Exploiting Temporal Structure*. Available online at: https://arxiv.org/abs/1502.08029v5. Accessed on 18th December, 2018.

Yeung, J. (2007) 'Audio Description in the Chinese World', in Dìaz-Cintas, J., Orero, P. and Remael, A. (eds.) *Media for All: Subtitling for the Deaf, Audio Description and Sign Language.* Amsterdam: Rodopi, pp. 231-244.

Yus, F. (2008) 'Inferring from Comics: a Multi-Stage Account.' *Quaderns de Filologia. Estudis de Comunicació,* 3, pp. 223-249.

Zabrocka, M. and Jankovska, A. (2016) 'How Co-Speech Gestures are Rendered in Audio Description: A Case Study', in Matamala, A. and Orero, P. (eds.) *Researching Audio Description, New Approaches*. London: Palgrave Macmillan, pp. 123-142.