Braun, Sabine; Starr, Kim; Laaksonen, Jorma

Comparing human and automated approaches to visual storytelling

Comparing human and automated approaches to visual storytelling

Sabine Braun & Kim Starr, University of Surrey, UK
Jorma Laaksonen, Aalto University, Finland

Abstract

This chapter focuses on the recent surge of interest in automating methods for describing audiovisual content whether for image search and retrieval, visual storytelling or in response to the rising demand for audio description following changes to regulatory frameworks. While computer vision communities have intensified research into the automatic generation of video descriptions (Bernardi *et al.*, 2016), the automation of still image captioning remains a challenge in terms of accuracy (Husain & Bober, 2016). Moving images pose additional challenges linked to temporality, including co-referencing (Rohrbach *et al.*, 2017) and other features of narrative continuity (Huang *et al.*, 2016). Machine-generated descriptions are currently less sophisticated than their human equivalents, and frequently incoherent or incorrect. By contrast, human descriptions are more elaborate and reliable but are expensive to produce. Nevertheless, they offer information about visual and auditory elements in audiovisual content that can be exploited for research into machine training. Based on our research conducted in the EU-funded MeMAD project, this chapter outlines a methodological approach for a systematic comparison of human and machine-generated video descriptions, drawing on corpus-based and discourse-based approaches, with a view to identifying key characteristics and patterns in both types of description, and exploiting human knowledge about video description for machine training.

1   Introduction

In recent years, there has been a noticeable surge of interest in methods for describing audiovisual content, whether for automatic image search and retrieval, for advanced audiovisual storytelling, or because of an increasing demand for audio description (AD) following changes in national and European broadcasting legislation to meet the needs of visually impaired

audiences. Approaches to creating AD for audiovisual content such as TV programmes and feature films are well established in some countries and can serve as models for other countries. However, the production of AD relies heavily on the specialised skills of audio describers and is therefore expensive. Attempts to automate visual content description, on the other hand, come with their own challenges. Although the computer vision and natural language processing communities have intensified research on automating image descriptions (Bernardi *et al.*, 2016), even the automatic description of *still* images remains challenging in terms of accuracy, completeness and robustness (Husain & Bober 2016). Descriptions of *moving* images and audiovisual content where visual and auditory channels combine for the purposes of audiovisual storytelling pose additional challenges linked to temporality, including co-referencing (Rohrbach *et al.*, 2017), and other features of narrative continuity (Huang *et al.*, 2016). Machine-generated descriptions are currently at best more semantically and syntactically naïve than their human equivalents; but they are often also incoherent or incorrect.

By contrast, human-made AD, which is the product of a highly creative process of intersemiotic translation (Braun 2016), provides one of the most elaborate and reliable types of content description currently available for (still and) moving images. Audio descriptions of audiovisual content such as TV programmes are not intended to be stand-alone texts. They are created and processed/consumed in conjunction with those elements of the audiovisual narrative that remain accessible for visually impaired audiences, i.e. the dialogue, narration, sound effects, music and song lyrics. However, when combined with these elements, AD is not only an effective means of making audiovisual content accessible but potentially also a source of information about visual, auditory and verbal elements in audiovisual narrative that can be exploited for research and machine training.

Against such a backdrop, this chapter reports on a study comprising a systematic comparison of human and machine-generated descriptions of audiovisual content, with the aim of identifying key characteristics of human descriptions that can inform and guide the development of (semi-)automated solutions. In line with the wider objectives of the project from which this study emanates, i.e. the EU-funded H2020 project 'Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy' (MeMAD), the aim is that these solutions can be applied to different contexts of use, especially content retrieval from broadcasting archives and content description for the benefit of sight-impaired people.

As pointed out above, the most relevant type of audiovisual content description for automation is AD. A further type of visual description that would benefit from AI intervention is content description for broadcasting archives. Anecdotal evidence suggests that this type of description is currently 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 re-use/re-sale of the content, i.e. the insertion of the content into another programme. Broadcasters therefore prioritise the description of content for which they own or have cleared or established the rights to support re-use internally or sale to other media companies. In contrast to AD, content descriptions for archival purposes are used in written form only, as an ancillary text to the audiovisual content, obviating the need for the descriptions to fit in audio hiatuses. Content descriptions also tend to be more 'literal' or factual than AD, especially AD for filmic drama and movies, which can at times be 'narrative' or figurative (Table 8.1). A model for machine-generated content description is therefore likely to be a more achievable goal in the shorter term than a model for generating elaborate audio descriptions. However, as pointed out above, AD can be used to derive guidance for automation. AD is also more widely and systematically accessible than content descriptions, which are an internal resource to broadcasters. Although the availability of AD varies in quantity, depth/detail and quality between countries and audiovisual genres, it is a rich source of information about the visual elements in audiovisual content and a relatively well studied source of insight into both how human understanding and human description of audiovisual content works. On balance, it therefore appears to be a suitable basis for modelling audiovisual comprehension and description.

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 8.1: Key features of different types of visual content description

As our first excursion into the topic of considering the merits of human-made AD for the modelling of its (semi-)automation, this chapter will begin by summarising key points from the study of human AD and human assimilation of audiovisual content (section 2) that are pertinent for the present study. This is followed by an overview of computer vision and machine-based approaches to audiovisual narrative and storytelling (section 3). Together these sections provide the foundations for the core of this chapter, i.e. a discussion of the methodological approach that we have adopted for this study (section 4). Although the focus of this chapter is on methodological considerations for this uncharted area of study, i.e. the elicitation and exploitation of human knowledge of audiovisual content description to advance automated solutions, we include a summary of observations from our pilot stage (section 5). We conclude by outlining what the piloting phase has highlighted and how this is shaping future steps (section 6).

2   Human understanding and description of audiovisual content

The study of (human) AD is mostly situated within the field of Translation Studies, where AD is characterised as a modality of intersemiotic translation, and more specifically as a practice of translating visual images or visual elements (and occasionally sounds that are incomprehensible without seeing the associated visuals) of audiovisual material into verbal descriptions. As we have shown in our research overview (Braun and Starr, in this volume [Introduction]), irrespective of specific outcomes of individual studies, research on AD highlights the complexity of this type of translation, including the complexity of information selection, prioritisation and verbalisation strategies; the advantages and drawbacks of different description styles; and the insight that whilst AD cannot be entirely objective, a degree of interpretation and subjectivity may lead to more successful AD. Given the relatively low level of sophistication that machine-generated descriptions of audiovisual content can currently achieve, the key characteristics of human-made audiovisual content description are likely to create challenges for machine-generated descriptions. However, the MeMAD project aim to advance the automation of audiovisual content description 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 of the remaining sections in this part of the chapter.

2.1 Cognitive-pragmatic frameworks of human 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 audiovisual content description, as they focus on explaining how we process monomodal verbal and/or multimodal (including audiovisual) content and retrieve the underlying story. Three particularly pertinent frameworks will be explored in more detail here, namely Mental Model Theory (Johnson-Laird 1983, 2006), Relevance Theory (Sperber & Wilson 1995) and Cognitive Narratology (Herman 2002, 2013).

Mental Model Theory (MMT) is essentially a theory of human reasoning. One of its basic postulates is that communication and understanding work on the basis of mental representations of what is being communicated, by virtue of combining perceptual input and prior knowledge. Mental models represent possibilities of how things could be in any 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 (Van Dijk & Kintsch 1983, 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. As the story unfolds, we normally settle on one of these in our interpretation of the textual cues (bottom-up processing) in light of the socio-cultural context of reception and common knowledge, including knowledge about places, activities and/or events (top-down processing). Mental modelling thus constitutes a process of hypothesis formation, confirmation and/or revision. Through its focus on the different sources of input, 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 is complementary in that it elaborates on some of the details of this process.

Relevance Theory (RT) focuses on the human ability to derive meaning through inferential processes. It 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 omitting information that can be retrieved from common knowledge) and that as recipients we develop them into full-blown semantic representations (propositions) as a basis for deriving the intended

meaning (Sperber & Wilson 1995). According to RT, we achieve this by retrieving the explicit and implicit assumptions (i.e. *explicatures* and *implicatures*) that a speaker is making. The retrieval of explicatures 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 temporal references or links between utterances), resulting in a basic level of utterance understanding. This is followed or complemented by the retrieval of implicatures to uncover a speaker's communicative message or intention.

RT asserts that these processes are highly inferential, drawing on common knowledge and cultural experience, that they are guided by the human tendency to maximise relevance (*Cognitive Principle of Relevance*) and our assumption that speakers/storytellers normally choose the optimally relevant way of communicating their intentions (*Communicative Principle of Relevance*) (Sperber & Wilson, 1995). In accordance with this, we stop processing an utterance as soon as we derive an interpretation that we find sufficiently relevant, regarding this interpretation as the optimally relevant interpretation as it provides the best balance between processing effort and effect. RT's 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.

The emergent field of Cognitive Narratology (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, Shank & Abelson 1977, Mandler 1984,). Schemata are thought to be part of our cognitive system. They include story schemata, i.e. 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). 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, and the way we recall and structure salient narrative during the act of story re-telling (Mandler & Johnson, 1977).

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. This has led them to emphasise the role of lexico-grammatical cues on the text surface ('cohesive ties') in the creation of textual coherence. Further research has demonstrated that coherence is in fact a much more complex concept (e.g., Blakemore 1992; De Beaugrande and Dressler, 1981; Brown and Yule, 1983; Bublitz and Lenk, 1999; Gernsbacher and Givón, 1995) and that the links needed to create continuity of meaning are supplied by text recipients whilst formal cohesion is neither a necessary nor a sufficient condition for coherence. However, a human storyteller will normally select appropriate means of expression to support the creation of temporal or causal coherence in the recipient's mind, including: temporal, causal and other link words; coreference chains; bridging inferences (Myers *et al.*, 2010); and motion verbs to create a sense of 'fictive motion' in a story (Talmy, 1983). Furthermore, focalisation (Bal and Lewin, 1983; Bal, 2009) as a function of both story and storyteller and often formulated by means of pronominalisation, creates an intermediate layer of narrative perspective (or 'bias') from which events are described and interpreted and which also impacts our understanding of story worlds.

2.2 Visual storytelling through the cognitive-pragmatic lens

Cognitive-pragmatic frameworks have traditionally focused on mono-modal and mono-lingual communication, but 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 (1995) do not discuss visual or multimodal discourse, but 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; Yus 2008; Forceville 2014). CN has been applied to both monomodal and multimodal storytelling, especially in filmic narrative (Herman 2002). Furthermore, there is a growing body of research using these frameworks to investigate multimodal translation (e.g.

Dicerto, 2018), audiovisual translation (e.g. Kovačič 1993; Desilla 2012, 2014) and audio description (Braun 2007, 2011, 2016; Fresno, 2014; Vercauteren & Remael, 2014).

One question investigated in this body of work is how, according to the cognitive-pragmatic models, meaning arises from multimodal and/or audiovisual content. Although Johnson-Laird (2006: 233) maintains that the cognitive processes involved in integrating cues from different sources into mental models are not well understood yet, there is consensus that in audiovisual co-narration, where different modes of expression are combined, their meanings are not simply added to each other but contextualise, specify and modify each other (Lemke, 2006).

Figure 8.1, taken from *Frida* (2002), illustrates this effect. The first dialogue turn in this extract ("And concentrate everybody") is invested with meaning by the accompanying visuals, which show Guillermo and his family getting ready to take a family photograph. Conversely, the two subsequent dialogue turns, which indicate that one family member—namely Frida—is missing, serve to frame the unfolding visual actions, namely Arianna getting up and walking off, and Frida entering the scene. The visual reactions of Frida's mother and sisters as Frida appears in a man's suit tell a story about their relationship with Frida. Together with the inferable knowledge that there are no male siblings in the family, these visual cues enable us to create a mental model about the family relationships that ultimately enables us to retrieve the ironic spin on Guillermo's penultimate utterance ("I always wanted a son") in this extract.

The AD for this extract captures many of these salient visual cues, providing a good example of how human audio describers enable visually impaired audiences to make sense of the dialogue/audio track. The description generally follows the main characters and describes their actions, using simple sentences in which the characters in focus are the agents ("Christina grins at Frida", "Mathilda sighs with exasperation"). Whilst most characters are only referred to by name, Adriana and Frida are assigned brief descriptions of their appearances ("plain featured" and "in a man's grey suit" respectively). Guillermo's joke ("I always wanted a son") explains why the detail of Frida's appearance, which is further reinforced by the reference to her trouser pocket, is crucial in the AD. The reason for describing Adriana's appearance is less obvious, but her appearance ("plain featured") contrasts with the more intriguing appearance of Frida, highlighting Frida's avant-gardist character. The other women's reactions to Frida's appearance ("Christina grins", "Mathilda sighs with exasperation") support this further.

8

1
Guillermo: And concentrate, everybody.
Christina: Wait. Where is Frida?

2
Mathilda: Adriana, go tell your sister to hurry up.

3
Plain-featured <u>Adriana</u> goes off to look for Frida, who appears in a man's grey suit, her black hair combed back.

4
<u>Christina</u> grins at Frida

5
<u>who</u> fixes a red rose into her lapel.

6
<u>Mathilda</u> sighs with exasperation.

7
<u>Guillermo's</u> eyes twinkle. he stands behind the camera, waiting to take the family photo.
Guillermo: I always wanted <u>a son.</u>

8
Guillermo: and, Mathilda, everyone, eyes to the camera, …and… [Click of camera]

9
In the black and white snap, Frida stands with her hand thrust into her <u>trouser</u> pocket.

Figure 8.1: Frida – taking a family photograph, old style[1]

At the same time the extract also illustrates that audio describers add aspects that are only inferable rather than being visible. This corroborates 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* (3), 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 (2). Similarly, the

---

[1] Images 1-8 were presented in full colour in the original film material, while frame 9 was rendered in black and white for narrative effect.

assumption that Christina's grinning is directed *at Frida* (4) is only inferable from the direction of Christina's gaze and our understanding of the preceding and subsequent shots (3 and 5), in which Frida enters 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 Figure 8.2, 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 can retrieve this from the narrator's voice that is reading out what she is writing and from further visual cues reinforcing the suicide note hypothesis (e.g. a note being left on the mantelpiece as the woman walks to a nearby river and begins to put 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 trembles as she fastens the buttons and ties the belt of her tweed coat.

4 5

Earlier she sits writing.

Figure 8.2: The Hours: Describing at different levels

As these two examples highlight, the complexity of human processing of audiovisual content means that *human description* of such 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 levels of granularity in AD may return useful descriptions, by exploiting the human ability to create mental models and derive meaningful inferences, may mitigate against some of the current problems with automatically producing elaborate video scene descriptions.

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

3   Computer vision and automated description of audiovisual content

Until recently, automatic audiovisual content description has consisted of techniques that detect visual and auditory elements from audiovisual content, and label them with pre-defined keywords or indexing concepts. Such keywords can be words derived from visual and aural categories and/or words recognised with a speech recogniser from the spoken utterances. This approach has severe limitations as, for example, accurate description of actions and the inherent properties of 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 more recent trend, large image and video datasets, 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 for each image or video object. Moreover, some image datasets, 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. These developments in the availability of training and testing data have opened up new avenues for devising more accurate and efficient methods for automatic visual data

description. 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., 2015)
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 and Dolan, 2011)
LSMDC	video	108536	108536	(Rohrbach et al., 2015)
MSR-VTT	video	6513	130260	(Xu et al., 2016)

Table 8.2: Image and video training datasets

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.*, 2015; He *et al.*, 2016). This representation is then fed to a recurrent neural network, typically a Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997), that decodes this visual encoding to an output sequence of words, a sentence or a caption that describes the audiovisual content.

Training the word sequence decoders for image and video content description has conventionally been based on minimising the discrepancy (cross-entropy) between the sentences generated by the model and the desired output. This approach is well-motivated theoretically, but does not aim to directly optimise any automatic performance measure used in practice such as BLEU, METEOR or CIDEr evaluation mechanisms. In order to improve the captioning performance with respect to these measures, researchers have started to use reinforcement learning (Ren *et al.*, 2017) in training the captioning models. This has led to

better results when measured by the automatically obtainable scores. Despite significant recent progress, current image and video description techniques are still unreliable, producing different textual descriptions for visually very similar contents.

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 *et al.*, 2018) has developed apace, most notably in relation to the description of object inter-relatedness within single frame images (Krishna *et al.*, 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 for the automated description of audiovisual content remains to produce a model for creating intuitive and coherent storytelling across multiple images read in sequence.

One of the challenges is that, while 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 Figure 8.1, from *Frida*, above), variations in scale or placement may 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.

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 *et al.*, 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, has 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.

Pairing automated audio event extraction and speaker diarization with image sequencing models could 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. Currently, however, automated approaches to video scene description remain largely confined to the description of individual frames within video scenes, reducing cohesion and coherence between descriptions of subsequent frames to a minimum. Initial analysis of automatically generated description needs to be adjusted to this state of affairs while remaining

open to a comparative analysis between more advanced combined sound-image machine-generated descriptions and their human-generated equivalents. This would ensure that the further development of automated solutions can be informed by insights into human approaches to description in the future. We believe that the methodological approach outlined in the next section is sufficiently flexible to accommodate this.

4 Methodological Approach

The initial phase of the present study has focused on three principal components. The first was the construction of a corpus of audiovisual materials consisting of human AD and original film dialogue (in English), the MeMAD Video Corpus (MVC) and the subsequent identification of short extracts within the corpus which lend themselves to human vs. machine generated description comparisons. The second component was the annotation of this audiovisual content in a manner which facilitates a comparative study featuring human and computer-generated video description, while the third component was a preliminary analysis of parallel datasets (human annotations, AD and a first iteration of experimental machine-generated video descriptions) to pilot the methodological design and initiate first improvements in automated descriptions. Each of these items was a key step in the preparation of more comprehensive comparative analyses between human-generated and machine-generated audiovisual content descriptions in the later phases of the study. This section elaborates on the first two components. Observations of the preliminary analysis are reported in section 5.

4.1 Materials

Selection of Materials

While audio described content is more readily available than other types of audiovisual content description (section 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. 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).

In addition, 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 in the context of our study. An example of the type of issues encountered was highlighted during the pilot phase, when the serial drama genre was explored as a potential source of audiovisual 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 AD. While this material contained useful examples of the kinds of narrative action which could theoretically inform human meaning-making in storytelling, 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* (2017), 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. 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 AD as the most comprehensive source of audiovisual data available for informing the creation of 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. As discussed in section 2.3, approaches to AD vary considerably in terms of style and granularity, and are ultimately subject to

the audio describer's personal filter and individual interpretation, life experience and intuition, all of which are tested against the benchmarks of redundancy and saliency. Perhaps not unsurprisingly therefore 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). This lack of standardisation naturally impacts objectivity, with considerable variation between describers in the way they choose to prioritise visual cues 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. As was highlighted in section 2.3, 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 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 in Table 8.3.

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 8.3: Advantages and disadvantages of AD for informing machine-generated descriptions

As highlighted above, 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 for analysing and comparing human and machine-generated descriptions of audiovisual content. However, our pilot work and the human models of communication reviewed above also make it clear that additional information would be necessary to compensate for the 'shortcomings' of AD in our context. 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 AD and machine-generated video descriptions, which we determined to be methodologically flawed. Instead, we decided to compile a corpus of audiovisual content with AD, which we would use as one source of information about the content, and to create different sets of annotations to complement the audio descriptive texts in this corpus.

Compiling the Audiovisual Corpus and Identifying 'Story Arcs'

Our primary experimental corpus, numbering forty-five 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).

Acknowledging the important role of story schemata in the 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 AD, 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 standardise 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	10 seconds – 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 8.4: 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.

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 (although this was not our research focus, as noted above), and a self-contained mini-plot in its own right. The duration of 'story arcs' was maintained between 10 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 8.3* below, and is taken from the film *Little Miss Sunshine*. At the beginning of the extract a dispute arises between a teenage boy and his family. The dispute is subsequently resolved by the intervention of a young family member. Screenshots of narratively key frames from the scene are shown 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.

Figure 8.3: 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 scene contains only minimal dialogue, allowing the AD to 'breathe' and deliver a relatively unhindered audio guide to the action. Although the majority of 'story arcs' selected for inclusion in our corpus contain dialogue in addition to AD, this example

21

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 video descriptions.

4.2   Data Processing and Annotation

Annotation Models and Levels

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) were considered as frameworks for annotating semantic and narrative content. However, the former applied tagging protocols that were considerably more granular than was required for our 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 it became apparent that a bespoke methodology would have to be developed.

Based on the theoretical frameworks of discourse processing / storytelling outlined in section 3 and in order to overcome the 'shortcomings' of AD in our context, as explained above, 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. 8.4, 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.
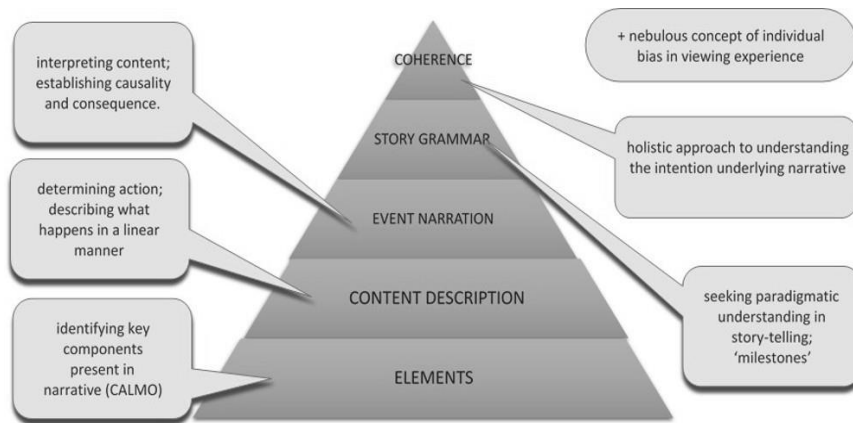
Figure 8.4: Accessing multimedia content – Levels of complexity

Our annotations have therefore been designed to address each of these levels of narrative immersion and can be described as follows:

(i) Key elements (KE): At the most fundamental level of meaning-making, our model assumes that viewers identify the building blocks of plot exposition, including main characters, actions, locations, the emotional temperature or mood of the piece, and salient objects, ('CALMO' in Fig. 8.4, above). 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. In accordance with this, we identified the KEs of each video scene as an entry point to the annotation and analysis process, using the following categories: *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* (e.g. happy, sad), *action-relevant object* (e.g. car, desk, bed) and optionally, *gestural/body language* (e.g. a shrug, a pointing finger). 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.

(ii) Content Description (CD): This annotation stream represents a 'ground truth' summary of the action taking place on screen. Constructed as a factual description while avoiding incursions into interpretation, CD captures the scene as it would be superficially perceived by the average audience member. In the Relevance Theory model of communication (Sperber & Wilson, 1995), CD corresponds to the level of *what is said*, i.e. the stage before any explicitly or implicitly communicated assumptions have been derived. Issues of causality and consequence

23

in relation to narrative actions were excluded as far as possible from this annotation level, these aspects being reserved for higher level annotations (below).

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 & Karuppali, 1980).

(iii) Event Narration (EN): This level captures a deeper pass of contextual cues, which we assume is conducted and applied by the viewer within the wider film context. EN reflects the viewer's attempt to establish relevance in relation to particular actions and construct context, which in turn informs understanding. In Relevance Theory, this corresponds, by and large, to deriving the *explicature*, i.e. the explicitly communicated assumptions. 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.

(iv) Story Grammar (SG): This 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. Referencing theoretical frameworks and the impact of Relevance Theory, this path to story resolution produces one or several implicatures derived from a summary of audio and visual cues. Accordingly, this layer of annotation corresponds to deriving the *implicature(s)*, i.e. the implicitly communicated assumptions, seen through the eyes of a sentient being endowed with pragmatic world knowledge.

The four layers will be used – along with the audio descriptions and film dialogue – as sources of data to evaluate comparable levels of sophistication in machine-generated video descriptions. The flexible nature of the annotation schema means that we are equipped to match any machine description iterations that are developed and generated in the MeMAD project. We can also adapt them in order to inform future computer vision models.

Transcription and Annotation Workflow

The initial transcription and 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, transcribing dialogue and AD as well as adding KE, CD and EN in one pass.

SG has not yet been added to our audiovisual corpus. 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).

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.

Validity of Human Annotations

Human beings make sense of the world from their own unique perspective. Following Mental Model Theory, 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 of 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. Future work exploring acceptable tolerance levels across related words will be required to resolve this issue.

Example of Annotation

To illustrate the annotation levels, this section presents an example of the annotation we have created to date, i.e. all levels outlined above with the exception of the SG layer. As is evident from the example shown below, which is the 'Boy in a field' scene from *Little Miss Sunshine* introduced in section 3.1, the KEs capture the *sine qua non* of the dramatic text. CD is 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.

The EN annotations record the 'why' for events occurring in the narrative and explicate cohesive links across the wider storyline.

The example also highlights the difference between the professional AD and our CD layer. The former strives to be complementary to the primary audio channels (dialogue and sound) and concise to keep pace with the film. As explained in section 3.1, the example contains very little dialogue, enabling a maximum amount of AD to be included, and yet, in the interest of striking a balance between AD and the other elements in the audio channel, the audio describer has left a range of visual cues for the audience to infer. By contrast, our CD annotation layer aims to give a systematic account of what can be seen on screen, irrespective of the time needed to read or perform the descriptions and of their placement in relation to other elements in the audio channel. Another point to note is that the AD does not introduce the characters. This is typical of our micro-narratives, where relevant information may have been introduced in the AD prior to the beginning of the video clip.

Key Elements:			
CHARACTER(S): a boy; a little girl			
ACTION(S): sitting, walking, hugging, climbing			
LOCATION(S): field (road)			
OBJECT(S): field, grass			
MOOD: sad			
OTHER: (gesture) hug			
Frame/Time codes	**Audio Description (AD) / Dialogue**	**Content Description (CD)**	**Event Narration (EN)**
1. 02:100994/01:07:19.760		Dwayne is sitting on the grass in a field, hugging his knees. He is sitting with his back to us.	Dwayne is upset.
2. 02:101125/01:07:25	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.	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.

3. 02:101650/01:07:46.000	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.	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.
4. 02:101875/01:07:55.000	She puts her arm around him and rests her head on his shoulder. His head turns slightly towards her.	Olive looks at Dwayne and then puts her arm around him, resting her head on his shoulder. Dwayne is trying not to cry.	
5. 02:102325/01:08:13.000	*Dwayne: I'm OK… let's go.*	Dwayne turns towards Olive. Dwayne reassures Olive that he is okay, and she looks at him and smiles.	Dwayne understands that Olive really cares about him and that she is genuinely upset for him.
6. 02:102475/01:08:19.000	Olive stands up and Dwayne gets to his feet and goes with her to the bottom of the slope.	Olive and Dwayne stand up and slowly walk towards the bottom of the slope.	
7. 02:102625/01:08:25.000	Olive starts to climb, putting out her hand for support. Dwayne lifts her up underneath her arms and carries her to the top of the slope.	Olive climbs the slope but she wobbles. Dwayne helps her by carrying her up. Olive seems to be smiling.	Dwayne helping his little sister appears to be a sign of him growing up and starting to care about his family. Olive looks proud for having helped her brother.

Table 8.5: Example of Annotation - Little Miss Sunshine ('Boy in a field')

5   Initial Observations

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 has taken 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. While reflecting the complex strategies for plot understanding and interpretation adopted by human audiences of film narrative, it enables us to compare the machine descriptions with each level of human description that was added to the corpus in the annotation process. The first example, i.e. the 'Boy in a field' scene repeated from above, but now including the first iteration of machine descriptions, is presented to illustrate the insights that can be drawn from comparing the machine descriptions with the different types and levels of human description.

Key Elements:
CHARACTER(S): A boy; a little girl.
ACTION(S): Sitting, **walking, hugging, climbing.**
LOCATION(S): Field **(road)**
OBJECT(S): Field, grass
MOOD: Sad
OTHER: (Gesture) Hug.

Frame/Time codes	Audio Description (AD) / *Dialogue*	Content Description (CD)	Event Narration (EN)	Machine Description (MD)
1. 02:100994/01:07:19.760		Dwayne is sitting on the grass in a field, hugging his knees. He is sitting with his back to us.	Dwayne is upset.	a **man** is **sitting** in a **field**
2. 02:101125/01:07:25	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.	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.	a **man** and a **woman** are talking to each other

3. 02:101650/01:07:46.000 	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.	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.	a group of people are singing and dancing
4. 02:101875/01:07:55.000 	She puts her arm around him and rests her head on his shoulder. His head turns slightly towards her.	Olive looks at Dwayne and then puts her arm around him, resting her head on his shoulder. Dwayne is trying not to cry.		a group of people are in a **field**
5. 02:102325/01:08:13.000 	*Dwayne: I'm OK... let's go.*	Dwayne turns towards Olive. Dwayne reassures Olive that he is okay, and she looks at him and smiles.	Dwayne understands that Olive really cares about him and that she is genuinely upset for him.	a **man** is running
6. 02:102475/01:08:19.000 	Olive stands up and Dwayne gets to his feet and goes with her to the bottom of the slope.	Olive and Dwayne stand up and slowly walk towards the bottom of the slope.		a **man** and a **woman** are **walking** in a **field**
7. 02:102625/01:08:25.000 	Olive starts to climb, putting out her hand for support. Dwayne lifts her up underneath her arms and carries her to the top of the slope.	Olive climbs the slope but she wobbles. Dwayne helps her by carrying her up. Olive seems to be smiling.	Dwayne helping his little sister appears to be a sign of him growing up and starting to care about his family. Olive looks proud for having helped her brother.	a **woman** is **walking** down the road

Table 8.6: Example of Analysis - Little Miss Sunshine ('Boy in a field')

The comparison of the machine descriptions with the key elements (KE) points to the first problem. With KEs covered in the machine descriptions highlighted in green it is evident that the machine descriptions are rather incomplete. In the above example, the computer algorithms miss several key actions such as walking and hugging and the mood of the scene. In some

frames, they also miss one of the two main characters. Furthermore, the repeated reference to the main characters as man and woman instead of referring to them more appropriately as boy and a girl is indicative for the lack of precision in current machine descriptions. A further noteworthy problem with regard to accuracy is the change from identifying the two characters as a man and a woman in segment 2 to identifying them incorrectly as a group of people in segments 3 or 4. This is difficult to explain other than by noting, as pointed out in section 3, that the production of very different textual descriptions of what is similar content to the human eye is a common phenomenon in current machine descriptions. In addition to the problems with character identification, several actions are also described incorrectly (e.g. in the final segment).

Another notable problem, which distinguishes the machine descriptions from all of the human descriptions and annotations, is the lack of relevance in several machine-generated descriptions. Although it is understood, as explained above, that the machine descriptions within one clip do not form a coherent narrative, as only individual frames are currently described, it is noteworthy that the computer vision algorithms often do not select the most salient actions even within individual frames. This is exemplified in the penultimate segment, where the MD reads "a man and a woman are walking in a field", suggesting an aimless action. Whilst the conclusion that they are in fact *returning* to the van after resolving the problem will only be possible when the algorithms become aware of how this frame is linked to previous frames/actions, a description to the effect that they are walking *towards the van*, which would create a more accurate and relevant description, may be achievable without sequential awareness.

The second example, drawn from *Saving Mr Banks* (2013), highlights further problems, especially the different ways in which the human descriptions and the machine descriptions approach cohesion and coherence across the descriptive segments, and the influence of the training data on the MD. In the selected scene, the main character, *Mary Poppins* author Pamela Travers, is angry with Walt Disney for filling her room with Disney branded toys in an attempt to seduce her into signing over the film rights for *Mary Poppins* to his corporation.

Key Elements:
CHARARCTER(S): woman
ACTION(S): walking, talking, carrying
LOCATION(S): bedroom
MOOD: angry
OBJECT(S): toys, sofa, cupboard
OTHER: gestures (sigh, hand dusting)

Frame/Time codes	Dialogue	Audio Description (AD)	Content Description (CD)	Machine Description (MD)
1. 00:00	P: Good riddance!		A woman with short hair wearing a brown suit closes the balcony doors.	A picture of a woman in a room
2. 00:03	P: Now ….	She dusts off her hands and steps out of her brown court shoes.	She kicks off her shoes …	
3. 00:05		She picks up a basket of Disney toys… and stuffs them into a wardrobe.	… and picks up a basket of stuffed toys from a sofa in the hotel room	
4. 00:10	P: Kids.		… carrying them to a cupboard	
5. 00:11	P: How old do they think I am?		… next to the hotel room door.	A room with a bed and pictures on the wall.
6. 00:16	P: Five years old or something?		She walks back across to the corner of the room…	
7. 00:18		She picks up Disney's 'Winnie the Pooh'.	… looks at another basket of toys …	A man standing in front of a mirror.
8. 00:19	P: Poor A.A.Milne…		… and picks up both the basket and Winnie the Pooh teddy bear.	A man holding a teddy bear in front of a mirror.

9. 00:22	P: Ghastly business!			She carries the baskets to the cupboard.	
10. 00:24 	P: Duck, dog, out!			She picks up Donald Duck and Pluto from the sofa…	A man is cutting a banana in a room.
11. 00:25 			She stuffs every single one of the toys into the wardrobe.	… and crams them into the top of the cupboard.	A woman is taking a picture of herself in the mirror.
12. 00:32				She slams the cupboard door shut.	
13. 00:34	P: [Sighs]				

Table 8.7: Example of Analysis – Saving Mr Banks

In this example, the AD is clearly less 'complete' than our CD, as the dialogue leaves little room for AD to be inserted. However, when processed along with the primary sound track (dialogue, non-speech sound), the AD provides the key information. The main character is correctly introduced (in this case, before the selected scene occurs in the film) and then correctly identified as the same person throughout the scene through 3rd person pronoun use. Furthermore, the audience can form an understanding (a mental model) of the character's main action, i.e. collecting stuffed toys from across the room and putting them in a cupboard, in a rather angry fashion, although some of the detail has to be inferred. For example, no explicit reference is made to the toys being scattered around the room but the AD segment "She picks up Disney's 'Winnie the Pooh'" (7), Pamela's emphatic comment "Duck, dog, out!" (10), the sound of objects being moved, jazzy music imitating steps, and the subsequent AD segment "She stuffs every single one of the toys into the wardrobe" (11) paint a picture of Pamela picking up a range of stuffed toy animals from across the entire room.

Our CD verbalises more of the visual detail and also demonstrates the process of character grounding more clearly than the AD: The character is introduced as new through the indefinite noun phrase "a woman" (1) and then repeatedly co-referenced through the 3rd person pronoun

("she"), creating a simple cohesive chain. By contrast, the MD fails with regard to character grounding and creating of a coherent sequence of action. It first introduces the main and only character in this scene correctly as a woman (1), but later—when only part of the character is visible—the MD wrongly identifies her several times as a man. (Whether this points to a bias in the training data is impossible to say from this one example, but it is an interesting question for further research.)

From the composition of the MD segments it is obvious again that they are unrelated, i.e. that they do not form a coherent whole. Each segment introduces a character or an object as new, using indefinite noun phrase constructions. This appears to be a reflection of the way the descriptions were composed in the training data, i.e. individual images described in a single sentence, by crowdsourced pieceworkers. Similarly, the first MD segment "a picture of a woman in a room" points to further problems with the training data: The captioners were instructed not to refer to the images they described as "pictures" or "images" but to focus on their content. However, this instruction seems to have been violated in several instances (Braun & Starr 2019).

Other aspects that stand out in the MD are the poor lexicon, the very restricted repertoire of syntactic structures and the striking errors in action identification (e.g. "cutting a banana", Table 8.7:10). The combination of the problems with the MD outlined in this section means that it would be difficult to create a coherent story from the MD. More broadly, the problems point to the differences between human audiovisual perception and machine perception, which are summarised in Table 8.8 below. The problems identified in this initial observation have informed the next steps of our analysis, which are outlined in the final section of this chapter.

Human perception	Machine perception
Moving images	Still images (single frames)
Character, action, location, mood recognition …	Object recognition
Narrative coherence	Neural networks
Relevance in meaning-making	Crowd-sourced captions
Life experience	Availability of training datasets

Table 8.8: Differences between human and machine perception of audiovisual content

6   Conclusions and next steps

The aim of this chapter was to outline an analytical procedure that supports a systematic comparison of human and machine-generated descriptions of audiovisual content with a view to using insights from this comparison to inform and advance the automation of visual or multimodal storytelling.

The theoretical models outlined at the beginning of this chapter make it clear that *human* visual or multimodal storytelling is a complex process with a range of uncertainties. The models explain why we draw different conclusions from the same premises and can give insight into why storytelling may be unsuccessful. Whilst, by emphasising the subjectivity of discourse 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 processing and storytelling also means that it has to date largely eschewed systematisation and formalisation. By extrapolation, the same applies to audiovisual content description including AD.

Similarly, progress in machine-generated descriptions of audiovisual content, i.e. descriptions of moving image sequences, has so far been modest, mainly because of a dearth of sufficiently large training and test data sets to assist machine learning. Models for creating coherent storytelling across multiple images read in sequence have yet to be developed.

Story grammar approaches, which first emerged in the late 1970s and have seen a recent surge in popularity appear to be a promising avenue for explaining and analysing (visual and multimodal) storytelling. As a schematised representation of events, processes and similar entities, story grammar lends itself to be formalised and may have a role to play in the development of computer models.

Given the current state of affairs, however, a more immediate step in our analysis will be a comparative lexical analysis of the human and machine-generated descriptions, seeking out differences in patterns of word use, informativeness values, omissions and misrepresentations. This analysis will be used to identify areas of interest, and examples subsequently selected for qualitative analysis on a case-by-case basis. 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. In

addition, we expect to extend the corpus-based lexical analyses to 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.

Further iterations of machine descriptions are expected to 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 and the addition of audio segmentation and diarisation techniques, i.e. extraction of sound features to measure impact, if any, on increasing inter-frame coherence. Combining audio and visual cues to infer continuity would contribute significantly to creating narrative coherence in automatic descriptions. If this approach proves tenable, we believe our human annotation and analytical methods are sufficiently agile to accommodate a comparative analysis between the combined sound-image machine-generated descriptions and their human-generated equivalents.

More broadly speaking, the increasingly complex association of ideas between frames presented in machine description outputs will 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 & Karuppali, 1980) added to our original annotations schemata, 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. 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, upon which the comprehension of narrative is contingent, and which subsume storyline expectations, episode schemata and plot units (Rumelhart, 1977; Lehnert, 1982), the sequencing of narrative and the interconnectivity between story components.

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.

Furthermore, the prospect highlighted in section 2.2 that different styles of audiovisual content 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.

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: 17 December 2018).

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.

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

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A. and Plank, B. (2016) 'Automatic description generation from images: A survey', *Journal of Artificial Intelligence Research*, 55(1), pp. 409-442.

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

Braun, S. (2007) 'Audio Description from a discourse perspective: a socially relevant framework for research and training', *Linguistica Antverpiensia, New Series-Themes in Translation Studie*,6, pp. 357-369, University Press Antwerp (UPA).

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: international journal on translation studies*, 28(2), pp. 302-313.

Braun, S. and Starr, K. (2019) [Forthcoming] 'Finding the Right Words: Investigating Machine-Generated Video Description Quality using a Human-Derived Corpus-based Approach', *Journal of Audiovisual Translation,* 2(2).

Blue Planet II (2017) BBC One Television. Available at: BBC iPlayer (Accessed: 11 December 2019).

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

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

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.

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, Oregon, USA, June 19. 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) *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: 11 December 2019).

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.

Desilla, L. (2014) 'Reading between the lines, seeing beyond the images: An empirical study on the comprehension of implicit film dialogue meaning across cultures', *The Translator*, 20(2), pp. 194-214.

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

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: 17 December 2018).

Gernsbacher, M. A. and Givón, T. (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. 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: 14 December 2018).

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

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

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.

Husain, S.S. and Bober, M. (2016) 'Improving large-scale image retrieval through robust aggregation of local descriptors', *IEEE transactions on pattern analysis and machine intelligence*, 39(9), pp.1783-1796.

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 www.audiodescription.co.uk/uploads/general/ itcguide_sds_audio_desc_word3.pdf. (Accessed: 18 December 2018).

Jiménez, C. and Seibel, C. (2012) 'Multisemiotic and multimodal corpus analysis in audio description: TRACCE', in *Audiovisual translation and media accessibility at the crossroads*, pp. 409-425. doi: https://doi.org/10.1163/9789401207812_022

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., and Zhang, B-T. (2018) *GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation.* Available at: https://arxiv.org/abs/1805.10973 (Accessed: 18 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.

Krishna, R., Zhu, Y.,Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L-J., Shamma, D., Bernstein, M.S., and 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.

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.

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.

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.

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.

Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J. and Lazebnik, S. (2015) 'Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models', *Proceedings of the IEEE international conference on computer vision (ICCV)*, Washington DC, USA, 7-13 December. pp. 2641-2649.

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: 18 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: 14 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 at: https://arxiv.org/abs/1704.01518 (Accessed: 20 December 2018).

Salway, A. (2007) 'A Corpus-based analysis of the language of 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 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 15-20. pp. 2556–2565.

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

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

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 at: https://arxiv.org/abs/1409.4842 (Accessed: 20 December 2018).

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

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.

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

Venugopalan, S., Rohrbach, 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 at: https://arxiv.org/abs/1505.00487 (Accessed: 18 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.

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 at: https://arxiv.org/abs/1609.08144v2 (Accessed: 18 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 at: https://arxiv.org/abs/1502.08029v5 (Accessed: 18 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.

Filmography

Frida (2002) Directed by Julie Taymor. [Feature film]. United States: Miramax films.

Little Miss Sunshine (2006) Directed by Jonathon Dayton and Valerie Faris. [Feature film]. USA: Fox Searchlight Pictures.

Saving Mr Banks (2013) Directed by John Lee Hancock. [Feature film]. UK: Walt Disney Studios Motion Pictures.

The Hours (2003) Directed by Stephen Daldry. [Feature film]. USA: Paramount Pictures.