

**MeMAD Deliverable***T5.2 Key Characteristics of Human and Machine Video Description*

Grant Agreement number	780069
Action Acronym	MeMAD
Action Title	Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy
Funding Scheme	H2020-ICT-2016-2017/H2020-ICT-2017-1
Version date of the Annex I against which the assessment will be made	8.5.2019
Start date of the project	1.1.2018
Due date of the deliverable	30 September 2019
Actual date of submission	Project manager will fill in
Lead beneficiary for the deliverable	University of Surrey
Dissemination level of the deliverable	Public

Action coordinator's scientific representative

Prof. Mikko Kurimo

AALTO –KORKEAKOULUSÄÄTIÖ, Aalto University School of Electrical Engineering, Department
of Signal Processing and Acoustics

mikko.kurimo@aalto.fi



MeMAD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069. This document has been produced by the MeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Authors in alphabetical order		
Name	Beneficiary	e-mail
Sabine Braun	University of Surrey	s.braun@surrey.ac.uk
Jaleh Delfani	University of Surrey	j.delfani@surrey.ac.uk
Kim Starr	University of Surrey	k.starr@surrey.ac.uk
Liisa Tiittula	University of Helsinki	liisa.tiittula@helsinki.fi

Internal QA			
Reviewer	Date of review	Comments	Date of revision
Sebastian Andersson	Oct 2019		29.10.19
Liisa Tiittula	Oct 2019		29.10.19
Lauri Saarikoski	Oct 2019		29.10.19

Abstract

This Deliverable is part of the MeMAD project's WP5, *Human processing in multimodal content description*, which explores human approaches to processing and describing audiovisual broadcast and media content (as a specific type of multimodal content), and compares them with machine-based approaches. In light of the advances and current limitations of machine-based approaches, and in line with the project's aim to advance this field, especially with regard to video scene description and audiovisual storytelling, it was decided that one of the project's work streams would focus on comparing machine-based and human methods for describing audiovisual content with the aim of identifying characteristic patterns of each method and informing the further development of machine-based algorithms.

This Deliverable describes the work carried out in Task 5.2, *Key Characteristics of Human and Machine Video Description*, which requires a comparative analysis using the crowdsourced captions derived from the training datasets, human-generated audio- and content descriptions, and machine-generated video descriptions derived from our computer model. The Deliverable begins with consideration of the training datasets employed to ready the machine ahead of processing the MeMAD Video Corpus (MVC) using the *DeepCaption* feature extraction model. Comparative analysis is then performed on the human descriptions and the first iteration of machine descriptions, as the sole source of images/captions upon which the MDs have been built.

A number of problems arising from both the training data and methodologies employed are discussed in relation to the first-iteration MD, including issues arising from the approach to crowdsourcing captions, actions taken to increase MD processing speeds and the ongoing difficulties associated with accurate object recognition and interconnectivity between multiple objects occurring in a single image. We conclude the first section of our report with a number of suggestions for improvement, both in relation to training data compilation and the delivery of algorithm and feature extraction for generating machine-derived video captioning.

In consideration of the second strand of our human vs. machine description research, we report on a case study conducted at Finnish national broadcaster, YLE, investigating the way archive editors search for and retrieve moving images for programme-making and re-sale. Interviews were conducted with the teams responsible for finding highly specific extracts from past broadcast productions for the purposes of commercial re-sale or in-house repurposing. In this case, the video captioning needs of the putative archive audience differ from those of the at home video consumer, with broader narrative concerns de-prioritised in favour of rapid retrieval via keywords and phrases.

Associated issues such as caption quality and end-user relevance are explored from both a practical and an ethical stance.

The report closes with a discussion of next steps to include: advances in multimodality; methods for promoting a diversification of the lexicon used in the machine-generated descriptions; and enhancing visual (and potentially audio) character tracking as a first step towards building sequential narrative.

Contents

Introduction.....	5
Part A	6
1 Introductory remarks	6
2 Audio Description and Content Description.....	7
3 Approaches to Analysing Video Captions.....	8
3.1 Creating the MeMAD500 Video Corpus ('MVC').....	9
3.2 Establishing narratively significant 'key elements'	10
3.3 Audio description capture	11
3.4 Creating the content descriptions.....	12
3.5 Production of captions for the MVC and training data	15
3.5.1 Video Captions.....	15
3.5.2 Training data.....	17
3.6 Corpus compilation and analysis.....	19
4 Corpus Comparison: Overview.....	22
5 Video Captions: Quality Assessment.....	24
5.1 Methodological Issues.....	24
5.2 Computer Vision Problems.....	29
5.3 Linguistic Considerations.....	37
6 Video Sequencing	45
Part B.....	48
1 Introductory remarks	48
2 Data Analysis	48
3 Content description at Yle.....	50
4 Analysis of Content Descriptions.....	55
5 Findings on the whole corpus	60
5.1 Structure.....	60
5.2 Level of granularity.....	61
5.3 Description of speech.....	63
5.4 Cohesion	64
5.5 Linguistic features	65
6 Summary	68
PART C.....	70
Conclusions and Recommendations	70

Introduction

In Deliverable 5.1, we reviewed the cognitive-pragmatic frameworks of human discourse modelling and storytelling and outlined the research design, data processing and annotation protocols applied in WP5, to demonstrate the work undertaken in this WP in the first year of the MeMAD Project. The present report reflects how the project has moved forward, shifting the focus to a comparative study of first-iteration machine descriptions produced by the MeMAD computer vision team, comparing them with human-generated descriptions of the same material. The MeMAD Video Corpus (MVC), created in the first phase of this study and outlined in D5.1, formed the basis of this work with descriptions and annotations of 500 film extracts being the subject of our analysis.

In **Part A**, we report on the comparative analysis of human descriptions of the MVC—i.e. content descriptions (CD) and audio descriptions (AD)—with the first iteration of machine-generated descriptions (MD) for the MVC using the *DeepCaption* feature extraction model. The analysis combines corpus-based and discourse-oriented approaches to identify the narrative elements that are characteristically selected for description and to explore how they are expressed linguistically. We identify similarities and differences in manually and automatically produced descriptions and evaluate the quality and usability of each type of description. To explore the differences further, we also discuss the process of creating crowdsourced captions which form the training datasets for the machine-based descriptions.

Part B reports on a separate study undertaken within the company archive service at Finnish National broadcaster, Yle. Moving the emphasis away from video description for consumer access (the focus in Part A) and in the direction of information search and retrieval in a commercial film archive, this section reports on current practices for generating metadata-type captions of video content for the purposes of re-use and re-sale. It is envisaged that understanding human methodologies and cataloguing behaviours in this way could indicate where improvements and efficiencies in practice might occur, as the precursor to greater standardisation and ultimately, semi-automation of marginal activities (i.e. those which are currently not cost-effective when performed by a human operative).

We conclude, in **Part C**, with recommendations and suggestions for future research paths both within the project and beyond. To this end, we consider: the validity of crowdsourced training data in the context of building models to perform complex, human-like tasks; issues of reliability in object identification and character recognition and possible approaches to resolving these; lack of cohesion and narrative sequencing in computer-generated captioning and perceived opportunities to explore artificial cohesion techniques though the application of linguistic and image-based strategies; the ethical and moral implications of endorsing semi-automation/automation as an ersatz form of video description which may meet the minimum requirements of regulatory frameworks and quotas, but is unlikely to deliver a service with any demonstrable consumer benefit in its current form.

Part A

1 Introductory remarks

We begin this Part of the Deliverable by recapping the nature of the MeMAD Video Corpus (MVC) and its purpose as a multimodal and narratively rich corpus of film material, moving on to discuss the two strands of human annotation used in the first iteration analysis, audio description (AD) and content description (CD). Owing to the disparate nature of these two texts, we discuss the comparative merits of human-generated audio descriptions (AD) and content descriptions (CD), outlining the rationale for working more closely with CD, which we regard as our descriptive audiovisual ‘ground truth’.

We then review the creation of these corpora (CD and AD) and our approach to annotation and analysis of the film material before shifting the focus to the machine-generated descriptions (MD) of the MVC, beginning with an overview of the state of the art of how video captions are produced and an exploration of the training datasets that were applied in the creation of the first-iteration MDs for the MVC corpus.

The MDs generated via the MeMAD computer vision model are then compared with ADs and CDs of the same source material using a corpus-based approach, in which we identify and explore grammatical, lexical and semantic patterns in the parallel corpora, first considering corpus-wide statistics such as type-token ratio, word frequency and keyness data, before investigating individual parts of speech (nouns, verbs, adjectives etc.).

This is followed by a quality assessment of the MDs. For this purpose, we group our observations into three principal categories, each of which impacts the quality of outputs: methodological issues, where problems are rooted in the nature of the training data; computer vision problems, which result from current limitations in object detection/identification; and linguistic problems, which are related to how the output of computer vision algorithms is rendered into natural language. This includes an exploration of relevant linguistic patterns in greater depth, engaging in the qualitative analysis of matters like lexical variation and granularity, semantic choices and, the impact of NLP and other factors on the production of linguistically cogent captions.

We conclude this part with a brief analysis of linguistic features that are particularly relevant in connection with video sequencing and coherence creation, especially pronominalization and the treatment of referents (e.g. as new vs. given). We note the differences between the various parallel corpora with regard to the use of these features.

2 Audio Description and Content Description

The initial expectation in the project was to harness human AD to inform the development of semi-automated solutions. A corpus-based approach was deemed appropriate, aimed at identifying patterns in human AD that are particularly relevant for the modelling of auto-generated descriptions. However, few AD corpora have been compiled to date, and even fewer are publicly available (Salway, 2007; Jimenez & Seibel, 2012; Rohrbach et al. 2015; Matamala 2019). Preparations to compile our own corpus showed that differences in stylistic factors, density and granularity of available AD meant much current TV production content is of limited use to the audio extraction processes originally envisaged in the project. For example, while TV drama contains useful descriptions of narrative action which give insight into human meaning-making in story-telling, the extent of the AD is constrained by quick-fire direction (multiple short scenes and rapid shot-changes) and a shortage of audio hiatuses, such that the corresponding AD is minimal and largely a vehicle for announcing changes of location (“in the pub...”) or introducing new characters (“Bernadette and Tiffany arrive”). Other TV genres also proved problematic. Documentaries, for example, generally lack a clear narrative within the AD, which serves the function of overlaying supplementary factual information where this is visually relayed. 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. 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. Lexically rich descriptions and contextualisation made feature-film AD a better candidate for inclusion in a corpus created specifically for our study. However, while AD has a perceived value in the context of informing machine-generated video descriptions, our pilot stage illustrated that extracting comprehensive visual information from AD can still prove problematic.

Irrespective of the differences between different audiovisual genres, in any material the absence of suitable hiatuses in the audio track, along with the ‘golden rule’ of AD that prohibits interruptions to the original sound track (Hyks, 2005), often limits the extent to which any supplementary visual information can be inserted into the source material. In the context of human comprehension this is not problematic. AD is not a stand-alone text; its purpose is to facilitate meaning-making in conjunction with the primary audio track containing dialogue, narration, sound effects, and musical scoring (Braun, 2011). It capitalises on the human ability to assimilate texts and sensory input by building mental models, establishing salience and relevance, and engaging skills of anticipation, inference and retrospective self-correction to retrieve the unsaid and the ultimately intended meaning (Braun, 2016; Fresno, Castellà & Soler-Vilageliu, 2016; Vandaele, 2012). This, in turn, like any other language mediation activity, encompasses an element of interpretation and subjectivity. Unsurprisingly, therefore, rule-based methodologies for arriving at audio described outputs have largely eluded AD producers and researchers (ITC, 2000; AENOR, 2005), as there is a lack of consensus between describers

about what should be included and omitted (Vercauteren, 2007: 139; Yeung, 2007:241; Ibanez, 2010:144) and considerable variation between describers in the lexical breadth with which they choose to describe the selected elements (Matamala, 2019).

Computer vision algorithms, by contrast, currently lack complex inferential capacity. Large-scale captioned image and moving image datasets of the type used for machine learning are not sufficiently numerous, sizeable or broad-reaching to bridge this gap. For example, while most available datasets (MS COCO, TGIF, Visual Genome, Rohrbach’s MPII-MD, Hollywood II) include still images or limited moving images, their application to training machines for the purposes of moving image description research is curtailed by the limited number of examples of each type of action or movement available. Whilst there are advances in parallel fields (e.g. task-driven facial recognition, emotion recognition, action detection etc.), the transferability of these different strands of research to narrative audiovisual content such as film is still a very challenging task.

What emerges from this is two-fold. On the one hand, existing **training datasets for machine learning are not entirely relevant to the description of narrative audiovisual content**. On the other hand, the **highly idiosyncratic and individualistic nature of human AD** suggests that it alone cannot provide sufficient data from which to elicit patterns that can inform and guide the automated production of human-like descriptions. In order to meet the requirements of the MeMAD project, namely, combining human knowledge of describing audiovisual content with machine learning and computer vision approaches, it became necessary to look elsewhere for human-produced descriptions of audiovisual content that can be used to identify patterns and strategies of human approaches. In short, the solution was to employ simpler human-produced ‘content descriptions’ (non-interpretative) which more closely matched the types of description the machine is currently capable of producing (non-interpretative, observational, object/action oriented). Of course, human-derived data inevitably includes a level of interpretation which introduces some element of idiosyncratic behaviour, as discussed above in relation to AD. However, our approach to creating content descriptions was to preserve a functionality that was as descriptive and objective as possible.

With regard to content descriptions, one set were created by the research team in English as a text sitting parallel to the AD and the machine description outputs, for the purposes of direct comparison (reported in the current Part A). In addition, a set of Finnish content descriptions supplied by Yle was analysed to explore authentic practices of making archive material accessible via search and retrieval practices (see Part B).

3 Approaches to Analysing Video Captions

Addressing the first task, as outlined above, i.e. that of analysing auto-generated video captions and comparing them with human-generated descriptions in order to understand

their structure and their current limitations led us to a **corpus-based approach** and the **compilation of human descriptive corpora that are comparable with machine description outputs**. For the reasons discussed above, this began with scrutiny of audio description texts. At first reckoning audio description appears the ideal candidate to fulfil the comparative brief as a linguistically and structurally sophisticated elaboration of the visual aspects of film material. Machine-generated video descriptions capture visual elements such as objects, characters, actions, locations and certain basic facial expressions, in a manner that is ostensibly similar to those selected by the human describer. However, the level of complexity in the narrative created by the audio describer far outweighs the lexically and syntactically naïve constructs currently produced by even the most advanced neural network model.

Furthermore, the human being draws on cognitive skills to infer what cannot be explicitly included in the AD due to time limitations which are likely to be beyond reach in the field of computer vision for the foreseeable future. As pointed out above, an alternative, plainer version of human description was therefore deemed to be an important stepping stone in creating a multimedia corpus which promotes direct linguistic comparison between professional audio descriptions, human-generated content descriptions and machine-generated descriptions. In addition, the type of audiovisual material to be used for this comparison needed to be considered carefully. As pointed out above, the genre of feature films offers the most complete and elaborate AD but is likely to be too complex for the current state of video captioning. This section explains our approach to creating datasets for the comparative analysis, i.e. our solution for the selection of audiovisual material, and the approaches to, and benefits of, creating different corpora of human descriptions, i.e. an AD corpus and a corpus with a ‘plainer’ content description.

3.1 Creating the MeMAD500 Video Corpus (‘MVC’)

As stated above, feature films were selected for our study because of their professional quality audio description and narratively challenging content. Since large-scale ‘off the shelf’ audio description corpora were not freely available, feature films which are already in the public domain and contain reliably accurate AD tracks, seemed a feasible alternative. Clearly, long-form and complex narrative of the type found in feature films is a giant leap for automated film captioning given the present state of the art, not least because concepts like sequencing and cohesion are absent. Nevertheless, a work-around for this problem was inspired by advances in automated visual storytelling (Huang et al., 2016) whereby short stories were devised by captioners using sets of five consecutive photos for the purposes of training the machine to orchestrate narrative. Our solution was to break down each of the feature films in our corpus into smaller, self-contained narrative units (somewhat similar to the short sequence photo experiment) with which, it was hypothesized, the machine might more successfully engage.

These took the form of stories-within-a-story (*micro-narratives*), containing clear, narratively significant beginning and end-points, and illustrating elements of crisis and resolution. However, the intention was that each ‘story-arc’ would be treated in isolation for the most part, without recourse to the greater insights available in the storyline beyond the micro-narratives themselves. In total, 501 extracts were studied from across a body of 44 feature length films, with each extract representing one brief micro-narrative (‘story arc’) of between 10 seconds and 2 minutes’ duration. Selection of an extract was dependent on there being a minimum of five separately identifiable images or actions across the duration, in order that the computer might detect visible change.

Mindful of the lack of sophistication in current machine-generated video descriptions, we selected examples of basic social interaction as the focus of 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	20 secs – 3 minutes	Limited duration story arcs should simplify sequence modelling
Storyline	Self-contained micro-narrative	e.g. initiating action/crisis, proposed solution, action based on solution, consequence, result
Objects	Unlimited	Although no limitation was put on the number of objects in an extract, only those objects regarded as key to the action were included in our annotations

Figure 1: Common features of video extracts

3.2 Establishing narratively significant ‘key elements’

As has been previously established (D5.1, p.35), audio description alone cannot supply the answers we seek in terms of a comprehensive and comparable text for training computer vision models to describe audiovisual material.

At the most basic level of meaning-making, as both consumers and creators of multimodal material, we are able to identify the fundamental building blocks of plot exposition. For the purposes of this study we chose to label these constituent parts ‘key elements’. They comprise

five essential markers which are universally present in traditionally structured narrative (post-modern and avant-garde storytelling being exceptions to this rule):

- main characters (e.g. man, woman, young girl, small boy)
- actions (e.g. sitting, walking, talking, eating)
- locations (e.g. at the office, in the kitchen, on a road)
- mood/ emotional temperature of the piece' (e.g. happy, sad, angry etc.)
- salient objects (e.g. car, desk, bed)

To this list, we added the *optional* 'gestural/body language' category (e.g. a shrug, a pointing finger) where called for in the film extract.

Establishing the nature of these important cues is generally the first task of the viewer, since without a gauge of mood, characterisation and the setting of narrative action, the viewer's inferential skills cannot be fully engaged. Whether or not these initial questions are answered instantly by reference to the film text, the viewer progresses to attempting an understanding of the action taking place, applying other kinds of multimedia cues to facilitate this process. These layers of meaning-making were discussed in detail in D5.1 (section 4.3) but essentially mark a non-linear progression from 'key elements' through a basic understanding of the on-screen action (our 'content descriptions'), to interpretation of actions by reference to the wider storyline ('event narration'), concluding with the application of story grammar principles to discern the shape of the narrative 'arc'. Viewer enlightenment is ultimately achieved through immersion in coherence seeking activities in order to extract inference and intention from the perspective of the storyteller.

As the first stage of multimedia accessibility, 'key elements' were explored not only as a means of deconstructing the mental modelling process, but were also extracted from the MVC for their potential to inform comparisons with metadata and other forms of moving image tagging, should this be automatically generated in the context of archive materials later in the project.

To summarise, the value of extracting 'key elements' as an entry point to the annotation and analysis process is that they are the *sine qua non* of dramatic texts. Although all of these elements may not be present at any single juncture, a combination of two or more at any given time will generally be critical to plot development and exposition and can therefore be regarded as narratively important.

3.3 Audio description capture

The audio descriptions were captured and transcribed as text from the audio descriptive track delivered in parallel with the selected film productions comprising the MeMAD Video Corpus (MVC). As such, this material was produced by professional audio describers and their scripts represent interjections typical of the kind advocated by film production companies (i.e.

dialogue-hiatus bound, narratively-driven, cognitively accessible). It was initially anticipated that such elaborate descriptions would provide information salient to the visual aspects of each film production against which the veracity and value of machine-derived descriptions created from the same source material might be assessed. However, not only is the process of arriving at relevant and timely audio descriptions highly complex as a cognitive and linguistic exercise it is, by its nature, also an incomplete text covering a very specific sub-group of visual elements required to aid (primarily) sight-impaired audiences.

In short, AD is applied to describe only those aspects of the film which the viewer cannot readily detect for themselves using the accompanying soundscape, whether dialogue, sound effects, non-verbal utterances or musical scoring. Visual cues for which simultaneous audio markers may be discovered either independently or in parallel with the on-screen action (e.g. dramatic music and the sound of a person screaming accompanying scenes of a burglary) and could therefore be regarded as redundant, are generally omitted from the AD. Such omissions represent a significant problem when considering AD in terms of a text through which to inform improvements to computer-generated video captions, given that the machine “sees” but does not simultaneously “hear” at present. For these reasons, it was concluded that AD did not provide the solution to training computers to deliver human-like video captions. AD does, however, represent a useful comparative text from which to determine the *narratively salient* visual cues from a human perspective in circumstances where these cannot be determined from the audio landscape. AD also contributes value in supplying data relating to the lexical characteristics of human description. Thus, as a professionally crafted corpus, movie AD can be said to comprise a high-quality body of material written in a style that is both lexically rich and narratively sophisticated. To this extent, the linguistic corpus derived from the AD track is reliable and considered (i.e. contains minimal errors either in comprehension of source materials or exposition in the AD output). The details of compiling the AD corpus are outlined in section 3.6).

3.4 Creating the content descriptions

Having determined that AD would not provide a one-stop-shop for sourcing linguistic material from which to extract comprehensive visual summarisations of film material, it was necessary to seek alternative annotations data in order to study human descriptive practices in comparison with machine video captioning. Our approach was inspired by our work with Finnish broadcaster Yle in the MeMAD consortium and by a consideration of archive retrieval approaches, meta data and ancillary texts (screenplays, scripts, programme guides). Archive retrieval within the broadcasting industry is founded in metadata and the tagging of video programming, and this practice is generally referred to as ‘content description’. Industry moving-image annotations are search-focused (personality-biased, relatively granular in nature, sales-oriented) and more prosaic than audio description, having less narrative interpretation and more overt labelling of key visual information.

As one strand of the project aims at enhancing automated description services, the creation of content descriptions for the MVC, designed to inform computer-led video search and retrieval, appeared to be a reasonably attainable goal. In order to safeguard objectivity as far as possible (bearing in mind that the points made about the subjectivity of AD apply to any form of human description/translation), the brief applied to building our human-generated ‘content descriptions’ corpus (CD) was to create a factual description of all discernible action occurring on screen while avoiding incursions into interpretation. Although the descriptions were kept brief, there was no need for them to fit around dialogue and other elements of the sound track. In practice, the standard applied to compiling content descriptions across the MVC was that the human annotator should identify actions and objects that are key to the narrative, and describe those elements in relation to each other and the micro-narrative within which they were situated, without reference to events or themes derived from outside the current film extract.

As a result, the CD corpus can be regarded as a further ‘ground truth’ against which machine descriptions, governed by similar limitations inherent within the automation model, might be critically evaluated. Predictably, lexical variation within the AD is 29.66% greater when measured against the CD corpus (using word-types, see *Figure 7*), which reflects the more filmic, descriptive remit prevailing in most AD guidelines. In the TGIF study, Li *et al.* (2016) compared AD (using the LSMDC dataset) and the human descriptions created in the process of captioning a set of animated GIFs. The LSMDC dataset was generated from commercial films and the descriptions were produced by professional descriptive video services; the TGIF dataset was created by online users and the captions were crowdsourced. The results revealed salient differences between the two datasets in terms of language complexity, visual/textual association and the scene segmentation. With regard to the language complexity, the professional describers used more complex and expressive phrases to make the videos more comprehensible for the visually impaired target audience whereas the crowdsourced captioners only described major visual content without using expressive language. In terms of visual/textual association, video descriptions often contain the contextual information that might not exist in a single video clip but can be grasped by humans from the video/film. By contrast, the animated GIFs lack any surrounding context. Following observation of this phenomenon, Li *et al.*, discovered that 20.7% of the sentences in LSMDC contain at least two pronouns, while in their TGIF dataset this number is only 7%. Another difference between the two datasets involved scene segmentation. Since the video clips in LSMDC are segmented through aligning speech recognition results to transcriptions, it is likely that some errors would occur in the process of sequence representation (either at the beginning or the end of the clip). This is not the case with the GIFs which are normally well segmented. The Li *et al.* (2016) study showed that 15% of the LSMDC clips and 5% of animated GIFs were not well segmented.

Figure 2 below shows the key elements for one of our clips (taken from *500 Days of Summer*), the dialogue transcript, the two types of human description we have used in our analysis to

data, and an example of a screenplay (in two versions – draft and final) to illustrate the differences. The inclusion of the screenplay also serves to illustrate why we did not pursue the potentially possible avenue of obtaining screenplays and using them in our corpus.

Clip #200115 (500 Days of Summer, 2009)					
Key Elements					
C- A man; a woman. A- Walking; browsing; talking, picking up (a record); showing (a record); smiling; opening (door); leaving (shop). L- Street; music store.			M- Sad. O- Records. OT- Rolling eyes.		
	Dialogue	Audio Description	Content Description	Script (2006 Draft)	Script (2008)
1		Later they are in a music store.	Tom and Summer are in a music store, browsing. Summer looks unhappy.	INT RECORD STORE – NIGHT Tom and a much more in control Summer walk down the aisles. He grabs one.	INT RECORD STORE – NIGHT Tom and a much more in control Summer walk down the aisles. He grabs one.
2	Tom: It pains me we live in a world where nobody has heard of <i>Spearmint</i> .			Tom: It pains me that we live in a world where no one's ever heard of <i>Spearmint</i> .	Tom: It pains me that we live in a world where no one's ever heard of <i>Spearmint</i> .
3	Summer: I've never heard of them.		Summer sounds annoyed.	Summer: I've never heard of them.	Summer: I've never heard of them.
4	Tom: I put them on the mixtape I made you. They're track one.		Tom is surprised.	Tom: And it's painful. Oh look.	Tom: They're on that disc I made you. (beat) They're Track 1.
5	Summer: Oh, yeah.		Summer looks uninterested.		Summer: Oh.
6		Summer nods, unconvincingly.	Tom rolls his eyes; he looks disappointed.		
7		Tom finds the <i>Ringo Starr</i> record.	Tom picks up a <i>Ringo Starr</i> record; he laughs and shows it to Summer. She smiles unimpressed.	He grabs a Ringo Starr album and shows it to her, just as we've seen on Page 7. She smiles and they continue on down the aisles.	Tom shakes that off, grabs a Ringo Starr album and shows it to her, just as we've seen in the beginning. She smiles and they continue on down the aisles.
8		Summer gives a tight smile and walks away from the record stand. Tom reaches out to take her hand, but she pulls away.	Summer walks away; Tom follows her and tries to hold her hand. Summer moves away and Tom looks sad.	In CU, Tom goes to hold Summer's hand. But something happens. It could be a total coincidence, but just as his hand approaches hers (in SLO-MO), she moves it away and keeps it at her side. Tom puts his hands in his pockets, unsure if there's something to read in that.	In CU, Tom goes to hold Summer's hand. But something happens. It could be a total coincidence, but just as his hand approaches hers (in SLO-MO), she moves it away and keeps it at her side. Tom puts his hands in his pockets, unsure if there's something to read in that.
9		With a disappointed sigh, Tom follows Summer out of the shop.			
10	[SFX WIND CHIME]		Summer opens the door and leaves the shop; Tom is right behind her.		
11	[SFX DOOR CLOSES]				

Figure 2: Annotations of clip 200115¹

¹ Screenplay: Neustadter & Weber (2008). *500 Days of Summer*.

The example shows that the Content Description provides a factual and continuous description of the main elements of the action, regardless of whether it ‘overlaps’ with the dialogue. It is intended for written use only. The Audio Description, by contrast, whilst also being factual and in the above example also largely focussed on what can be seen, does not provide a continuous description, as the AD segments alternate with the dialogue throughout the sequence. The screenplay is shown here in different versions, which are both available online. However, screenplays are difficult to obtain (as opposed to film scripts made by fans and normally containing no more than the film dialogue). In addition, they are not necessarily correct and/or difficult to process. For example, the 2008 version of *500 Days of Summer* is available only as a non-searchable pdf. Although it is closer to the actual film dialogue than the 2006 draft, some discrepancies remain. More important in our context, the descriptions are not necessarily complete (see e.g. sections 5 and 6) and they do not always present physical descriptions of what can be seen (e.g. in section 7 “Tom shakes that off” and throughout section 8). They also sometimes contain references to the script itself (see section 7), and not all descriptions are correct (e.g. the reference to SLO-MO in section 8).

In relation to our further explorations, i.e. explorations relating to the structure of the micro narratives (‘Story Grammars’), it is also noteworthy that this information is not normally indicated in the script. Only in exceptional cases is a detailed analysis of the film available, which may help in story grammar analysis. In the case of *500 Days of Summer*, for example, *script reader pro*, a website teaching screenplay writing² deconstructs the film’s screenplay into the characteristic seven-sequence, three-act structure, whereby each act is shown to be constructed of several steps (inciting incident – call to action – midpoint – big event/turning point – denouement).

3.5 Production of captions for the MVC and training data

3.5.1 Video Captions

The film clips in the MVC corpus were sub-divided into three tranches, horizontally (first-third of film extracts from each movie belonging to one tranche; second-third of film extracts from each movie belonging to the second tranche, etc.). By processing the film material in this way, we had planned to use the first tranche clips to produce first-iteration descriptions, and reserve the second and third tranches for later machine iterations. The concept underlying this corpus splitting exercise was that results produced from later iterations of the machine description algorithm might potentially become corrupted by the film material having previously been exposed to machine processing as test data. However, the dangers of test data serving dual-purpose as training data in this manner were considered to be negligible. For this reason, it was decided to process all clips in the film corpus to produce the first

2006 version (draft): <http://www.cinefile.biz/script/500daysofsummer.pdf>.

2008 version:

<https://static1.squarespace.com/static/5a1c2452268b96d901cd3471/t/5b987b06cd8366dd19611a09/1536719631052/500DaysOfSummer.pdf> via <https://screenplayed.film/scriptlibrary/500-days-of-summer-2009>

² <https://www.scriptreaderpro.com/500-days-of-summer-screenplay/>

iteration machine description captions. This also had the advantage of providing more comprehensive data with which to perform our analysis. The same material will be re-processed at later points to generate further iterations and case studies of selected phenomena. In the latter, we will use one or more of the sub-corpora, as appropriate.

Hence, the first-iteration corpus of captions (machine descriptions) was created by applying Aalto's *DeepCaption* model (Sjöberg *et al.*, 2018) and using two large-scale open access datasets for visual object recognition as training data, i.e. MS COCO (Lin *et al.*, 2015) and TGIF (Li *et al.*, 2016) – a combination referenced as the 'dc-a3' model. In addition to the training data, Aalto's *DeepCaption* software exploited the combined aspects of RNN for object identification and CNN for caption generation.

Multiple captions were created for each of the 501 MVC clips, with one caption being generated by the machine at each computer-detected shot change. This means that the computer model is not applied to moving images *per se*, but operates on the basis of describing a single frame at a time (in our iteration, the middle frame of a shot), each of which is considered in isolation from the remaining imagery and any associated context. The quality of the resulting video captions is largely dependent on the quality of the image descriptions contained in the training data and model feature extraction, since the captions are sourced from these datasets.

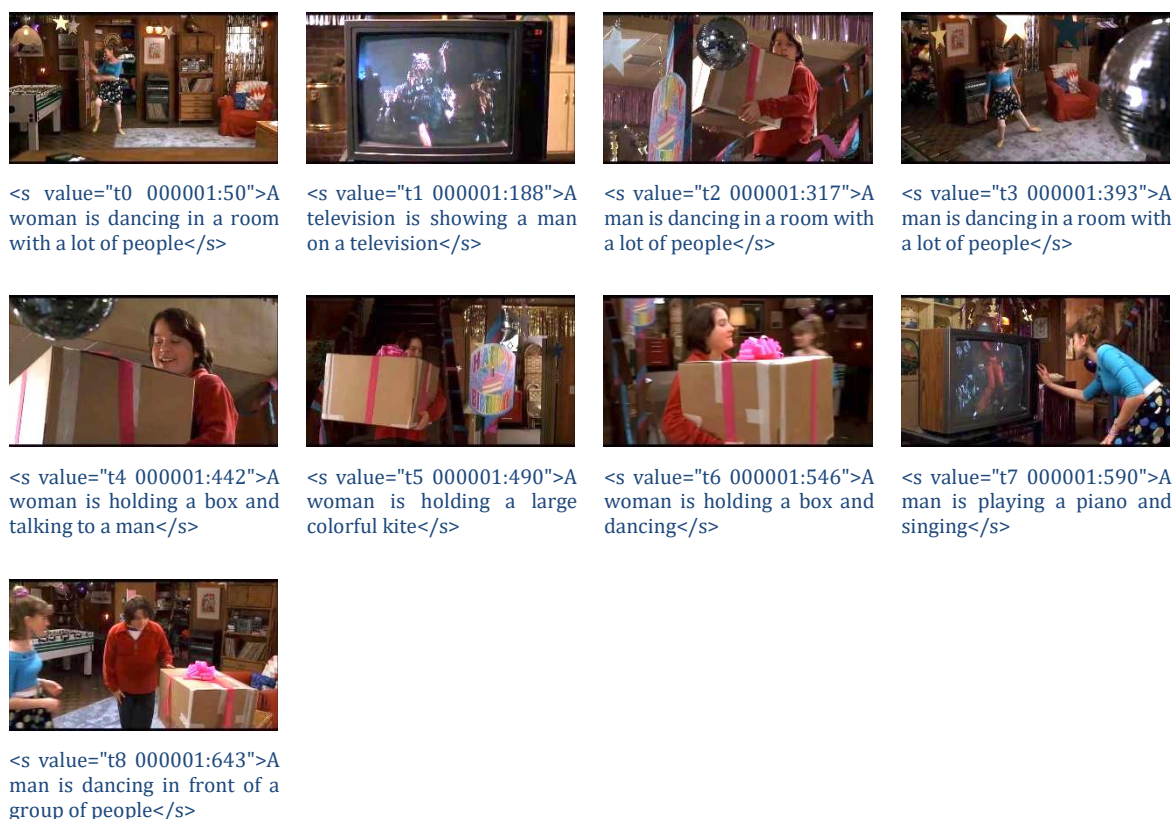


Figure 3: Example of first iteration machine description (Clip 00001)

A number of key characteristics, common across the corpus, can be observed in the above MD captioning sample (single film extract), namely:

- (i) the correct identification of certain items: *woman* (in caption 1), *TV*, *box*;
- (ii) other items are present in the description, but incorrectly identified: *man* (instead of *woman*), *kite*, *a room with a lot of people*, *dancing* (in two instances);
- (iii) a number of objects are not identified: one of more of the characters in some frames;
- (iv) narrative coherence is lacking because, as explained above, the current model selects individual frames only and is programmed to caption each independently of the next;
- (v) while syntactic structure in the MD favours animate subjects, mostly ‘A man is ...’ or ‘A woman is ...’ (e.g. ‘A woman is sitting on a couch and talking’, ‘A man is dancing in front of a microphone’), the proportion of captions starting with an inanimate object is approximately the same across the three sub-corpora (MD, CD, AD; see Fig. 4);
- (vi) the difference between MD and CD/AD in this regard is that the latter both make use of human inferencing to convert ‘a door’ in one shot, to ‘the (already referenced) door’ in subsequent shots. The machine model is not yet designed to connect images or conceptualise a door in the same way as a human, and thus treats every occurrence of the same door as ‘a door’ (see 5.3 below).

Corpus	Total number of captions	Number of captions starting with inanimate objects
MD	7,067	238
CD	4,892	138
AD	2,524	58

Figure 4: The number of captions starting with inanimate objects in the three corpora

3.5.2 Training data

MS COCO comprises 2.5 million instances of objects in 328k images harvested from the social media website *Flickr*. Each image was annotated with one-sentence captions by five individual operatives (Chen *et al.* 2015), as shown in Figure 5. **TGIF** consists of 100k short sequence animated images (GIFs) drawn from *Tumblr* and annotated with 120k natural language sentences. Both MS COCO and TGIF were compiled by harnessing the power of crowdsourcing (Amazon Mechanical Turk, AMT) to produce the annotations.

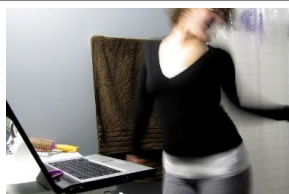



	<p>107963 http://cocodataset.org/#explore?id=107963 a girl is dancing in the bathroom to music on her lap top. a woman dances to the music playing on her computer a young woman is dancing in front of a laptop on a desk. a woman that is standing in front of a laptop. a young female is dancing in her bathroom.</p>
	<p>330053 http://cocodataset.org/#explore?id=330053 a lady observing a woman carrying two large bags and a man doing karate a woman is walking down the sidewalk carrying two large bags and a man is one the sidewalk dancing. a man roller blades down a city street. a man dancing on a sidewalk near a fire hydrant. the man is dancing on the sidewalk in front of everyone.</p>
	<p>477156 http://cocodataset.org/#explore?id=477156 a living room has a large box placed in the middle. a living room with a box for a large screen tv sitting in the middle of it. a large box sits on the floor in between the couch and coffee table. a living room with a very large unopened box located in front of the coach. a large brown box in front of a burgundy couch.</p>
	<p>338317 http://cocodataset.org/#explore?id=338317 there is a lot of foot traffic on this street during the day. people walking down a sidewalk near a road and a building. a street with various people walking by a building. there are people that are walking on the street an image of a person walking down the street on her phone</p>

Figure 5: Examples of (human) captioned image from MS COCO


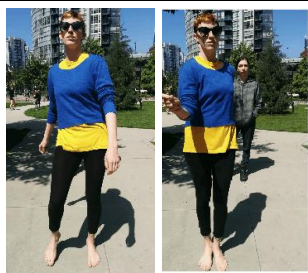

	<p>GIF# 002484 a woman is dancing along to what is showing on the television screen</p>
	<p>GIF# 000789 a woman in a blue and yellow shirt is dancing outside</p>
	<p>GIF# 000974 a group of women dressed in white are dancing</p>

Figure 6: Examples of (human) captioned images from TGIF

The way in which the captions in the training datasets were used to produce the captions for the MVC is difficult to identify in detail but as a general tendency, the MVC captions do not normally contain entire sentences/captions from the training data; they combine fragments of different captions taken from one or both datasets. For example, the first caption in MVC clip #000001 (*Figure 3* above), ‘A woman is dancing in a room with a lot of people’, does not appear verbatim in MS COCO. However, as *Figures 5 and 6* illustrate, images in MS COCO and TGIF show several people dancing in different settings.

3.6 Corpus compilation and analysis

Our strategy has been to commence with a quantitative analysis of human and machine-generated descriptions using corpus linguistics tools and techniques, and focusing on lexicogrammatical phenomena (see *section 4* below). In order to reach this point, the annotations representing our ‘ground truth’ – i.e. content descriptions and key elements (characters, actions, locations, mood and gestures) – as well as the transcripts of the professional audio description and film dialogue, all of which had been constructed by a team of annotators in year 1, were transformed into a set of **parallel text corpora** and **aligned with the video clips** to which they refer. The additional descriptions that we created for the purposes of narrative sequencing, which we termed ‘event narration’, will be used at a later stage – i.e. a stage when the machine descriptions have evolved more – to plot storylines in conjunction with elements of story grammar. These event narrations consist of contextualised commentary on the significance of narrative events to the story-telling arc, based on human inference and interpretation.

As previously noted, the data preparation and processing focused on converting the different layers of annotations of our 500 video extracts (‘*micro-narratives*’) from 45 films into parallel corpora, aligned with each other and with the film extracts: Audio description and dialogue were transcribed from the original screenplay; a summary of the ‘key elements’ present in each extract was supplied as list of key words denoting respectively, characters, actions, location, mood, objects and gestures. Content descriptions created by the annotators represented a brief summary of the narrative action as it occurred in each extract (‘say what you see’). Since AD is, by its nature, an incomplete rendition of mainly visual markers, we consider content descriptions to be a more reliable ‘ground truth’ against which the validity of the machine descriptions can more equitably be measured.

After completion by three independent transcribers/annotators, the textual annotations were passed to the main researcher for review to ensure consistency of descriptive/narrative style and in levels of granularity. The texts were normalised for consistency in rendering aspects such as non-verbal utterances, abbreviated text, numeration, narrator interjections, sound effects and other non-verbal audio elements. Basic information about the resulting corpora is shown in *Figure 7*.

	Human Content Description	Human Audio Description	Machine Description Iteration 1
Word tokens	43,829	25,039	70,315
Types	3,061	3,969	580
Type-Token Ratio	0.067	0.158	0.008
Lemmas	2,356	3,108	518
Sentences	4,892	2,524	7,067

Figure 7: Basic corpus information

The final step in data processing was to apply XML/TEI tags to encode the main characteristics of the texts (clip IDs, time codes, sound effects etc.). The same principles were later applied to the machine-generated descriptions. [<p> <s> and <align> are elements of SketchEngine notation and the remainder are derived from TEI as the linguistic standard for tagging]

Audio description	Content description
<p><align><clip number="000501" time="00:08:02 00:08:26">	<p><align><clip number="000501" time="00:08:02 00:08:26">
<s>An accident has brought traffic to a standstill in a busy city street. <s>	<s>A pan shot rises from static car to view the street scene from above. A stream of cars are caught in a traffic jam. Cuts to Bruce in his car, with a shot of the rear-view mirror from which hang a string of beads. <s>
<s>Bruce flicks the beads.<s>	<s><s>
<s>He shakes his head incredulously.<s>	<s><s>
<s><s>	<s>Bruce is sitting in a silver grey car, looking irritated. He flicks the beads.<s>
<s><s>	<s>He rotates the steering wheel back and forth in annoyance.<s>
<s><s>	<s><s>
<s>He pretends to drive maniacally. <s>	<s>He holds onto the steering wheel and pretends to drive crazily. <s>
<s><sound type="BLEEPER"/> His bleeper sounds.<s>	<s>Bruce looks at his bleeper and then replaces it in his trouser pocket.<s>
<s><s>	<s>Bruce shouts out loud, although he is alone in the car.<s>
</clip></align></p>	</clip></align></p>

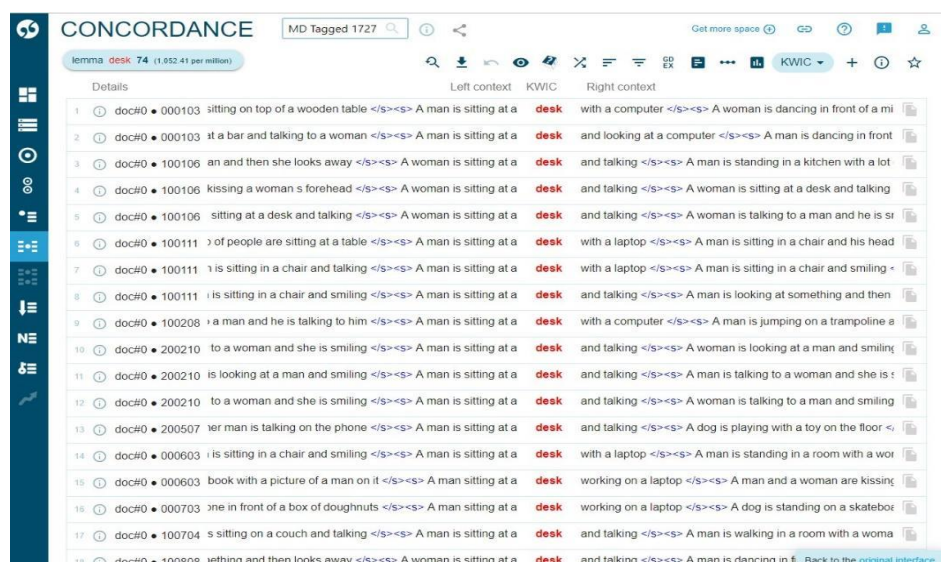
Figure 8: Audio and content description coding

The examples show once more how the AD provides summary descriptions (due to time constraints) rather than specific information about physical objects, actions etc. (see e.g. segment 1). The content description is more detailed.

During the multi-layered approach to corpus creation, a number of software packages for text/corpus and multimodal analysis were tested. The aim was to find a package that would enable us to align multiple parallel corpora simultaneously with the audiovisual content, to allow for direct comparisons to be drawn. However, the identification of suitable software tools turned out to be one of the key challenges. None of the multimodal software packages tested to date (MaxQDA, GATE, Elan amongst others) met our exacting requirements for multimodal analysis fully but work on this continues and will feed into more fine-grained (qualitative) analysis of the multimodal data.

ELAN seemed to be the obvious choice in the first instance, as machine-generated captions produced by the computer vision team can be easily read in the software, and files created may be readily exported in XML format. However, besides manual manipulation of data, there appeared to be no solution to time-aligning each of the human-generated corpora for direct comparison. We also explored the open-source software GATE (<https://gate.ac.uk/>), a computational linguistic programme designed to handle human language media using pipeline processing tasks. Although useful for information extraction and tagging, its limitations, particularly the lack of a flexible interface for processing moving images and linking these with corresponding corpora, proved an insurmountable barrier to application in the processing of multi-stream multimodal data.

As a solution, the textual data were ingested into an established corpus analysis tool (*Sketch Engine*), which supports alignment of multiple parallel corpora and export in XML format. The video clips are linked to the relevant corpus segments via the encoded clip IDs.



Details	Left context	KWIC	Right context
1 doc#0 • 000103	sitting on top of a wooden table </s><s> A man is sitting at a	desk	with a computer </s><s> A woman is dancing in front of a mi
2 doc#0 • 000103	at a bar and talking to a woman </s><s> A man is sitting at a	desk	and looking at a computer </s><s> A man is dancing in front
3 doc#0 • 100106	an and then she looks away </s><s> A woman is sitting at a	desk	and talking </s><s> A man is standing in a kitchen with a lot
4 doc#0 • 100106	kissing a woman s forehead </s><s> A woman is sitting at a	desk	and talking </s><s> A woman is sitting at a desk and talking
5 doc#0 • 100106	sitting at a desk and talking </s><s> A woman is sitting at a	desk	and talking </s><s> A woman is talking to a man and he is sr
6 doc#0 • 100111	of people are sitting at a table </s><s> A man is sitting at a	desk	with a laptop </s><s> A man is sitting in a chair and his head
7 doc#0 • 100111	is sitting in a chair and talking </s><s> A man is sitting at a	desk	with a laptop </s><s> A man is sitting in a chair and smiling
8 doc#0 • 100111	is sitting in a chair and smiling </s><s> A man is sitting at a	desk	and talking </s><s> A man is looking at something and then
9 doc#0 • 100208	a man and he is talking to him </s><s> A man is sitting at a	desk	with a computer </s><s> A man is jumping on a trampoline a
10 doc#0 • 200210	to a woman and she is smiling </s><s> A man is sitting at a	desk	and talking </s><s> A woman is looking at a man and smiling
11 doc#0 • 200210	is looking at a man and smiling </s><s> A man is sitting at a	desk	and talking </s><s> A man is talking to a woman and she is
12 doc#0 • 200210	to a woman and she is smiling </s><s> A man is sitting at a	desk	and talking </s><s> A woman is talking to a man and smiling
13 doc#0 • 200507	er man is talking on the phone </s><s> A man is sitting at a	desk	and talking </s><s> A dog is playing with a toy on the floor
14 doc#0 • 000603	is sitting in a chair and smiling </s><s> A man is sitting at a	desk	with a laptop </s><s> A man is standing in a room with a wor
15 doc#0 • 000603	book with a picture of a man on it </s><s> A man sitting at a	desk	working on a laptop </s><s> A man and a woman are kissing
16 doc#0 • 000703	ne in front of a box of doughnuts </s><s> A man sitting at a	desk	working on a laptop </s><s> A dog is standing on a skateboar
17 doc#0 • 100704	s sitting on a couch and talking </s><s> A man is sitting at a	desk	and talking </s><s> A man is walking in a room with a woma
18 doc#0 • 100808	ething and then looks away </s><s> A woman is sitting at a	desk	and talking </s><s> A man is dancing in

Figure 9: Sketch Engine, example concordance in machine captions dataset

4 Corpus Comparison: Overview

Comparison of the three key corpora (machine descriptions, human-created content descriptions and audio descriptions) illustrates the fundamental differences between video descriptions produced as a result of basic machine learning, and those derived from human interaction with the same multimodal materials. Before turning to these, it should be noted that in terms of overall corpus size, the AD corpus is – as expected – smaller than the CD corpus, given the purpose and brief of the content descriptions (see above). The MD corpus is the largest, although the size is entirely arbitrary since the frequency/points at which the machine produces a caption can be adjusted by time interval (e.g. every 3 or 10 seconds), frame count or shot-change detection. As explained above, in our first iteration a caption was generated for the middle frame of each shot.

Category	MD Types	Tokens	CD Types	Tokens	AD Types	Tokens
all words	580	70,315	3,061	43,829	3,969	25,039
type-token ratio (TTR)	0.008		0.067		0.158	
nouns	363	18,160	1,482	13,403	1,862	7,291
verbs	88	18,964	531	9,576	726	4,458
adjectives	39	460	297	1,448	490	1,221
adverbs	7	1,783	179	1,917	250	1,097
conjunctions	2	4,498	5	2,077	5	985
pronouns	14	1,938	21	3,477	21	2,888
prepositions	22	8,500	60	5,232	52	3,300

Figure 10: Corpus information and comparison

The number of unique words (*types*) represented in the MD corpus is considerably smaller – even in absolute terms, despite the larger size of the MD corpus – than that present in both of the human description modalities (MD: 560; CD: 2,941; AD: 3,951), illustrating at a glance the lexical poverty in the automated output. A similar pattern can be observed in relation to verbs (MD: 88; CD: 531; AD: 726) and adjectives (MD: 39; CD: 297; AD: 490).

In each case, the percentage of unique words appearing in the machine corpus as a percentage of the CD equivalents are: all words (19.72); verbs (16.57); adjectives (13.13). Whilst the same comparison in relation to uniqueness in the MD vs. AD corpus produces the following scores (%): words (14.68); verbs (12.12); adjectives (7.96).

The type-token ratio (TTR) of the three corpora (MD 0.008, CD 0.067, AD 0.158) supports this observation. As can perhaps be expected, the professionally created audio descriptions have the highest TTR, meaning that the lexical variation in this corpus is greater than in the other two. However, the TTR of the CD corpus is in the same order, whilst the TTR of the MD corpus is 20 times lower than that of the AD corpus and 8 times lower than that of the CD corpus. For comparison, TIWO, the AD corpus built by Salway (2007) based on AD of different TV genres, registers a TTR score of 0.044, and the LSMDC corpus (Rohrbach et al., 2015), which contains 180 films with professional AD, has a TTR of 0.021.³

These descriptive statistics paint an unequivocal picture of the overall shape and parameters of the machine corpus, which clearly falls short of human descriptions in all areas of lexical diversification. Indeed, not only is the size of the MD lexicon an average 17.2% of that created by human operatives (across AD and CD modalities), but adjectives comprise 10.9% of the CD corpus and 12.4% of the AD corpus, yet only 6.7% of the machine corpus. It is perhaps not surprising that the human operative annotations deliver a description that is more creative, imaginative and entertainment-led than the machine currently produces, although this imbalance might potentially be partially rectified in future machine iterations by changes to computer vision feature extraction.

Notably, adverbs are largely absent in the MD corpus (word tokens: 1783; type-tokens:7) with a high number of word tokens generated by only two types: ‘then’ (1,091) and ‘away’ (680). The derivation of ‘then’ can be traced back to an anomaly in the training data which resulted in split-screen images being captioned by crowdsourced operatives as if they were two images, conjoined with the phrase ‘and then’ (see 5.1). Regarding the adverb ‘away’, of the 680 word tokens found in the MD corpus, 601 are collocates of the verb ‘look’. The remaining five adverbs have a frequency of four, or less, in the MD corpus.

This quantitative overview serves to illustrate the differences between the corpora. Further insights come from our comparative qualitative analysis of the data for the purposes of identifying characteristic features and pattern deviations between machine- or human-led approaches. These insights will be outlined in the next section, which focusses on an assessment of the current quality of machine-generated descriptions.

³ Due to the much larger size of the TIWO and the LSMDC corpus (300k and 1M words respectively), the TTR of these corpora is only a rough indicator, as it is natural for the TTR to decrease with corpus size. In the TIWO, the different TV genres from which the audiovisual content for this corpus was drawn, may also have had an impact on the TTR. The LSMDC corpus contains 1,080,922 word tokens, 22,975 types, 16,507 lemmas, and 108,536 captions

5 Video Captions: Quality Assessment

Our initial quantitative analyses of the machine descriptions, as exemplified in *Figure 3*, show that at present, these descriptions hardly give insight into the essence of many of our micro-narratives. On the face of it, the computer algorithms often miss or mis-identify one or both of the main characters, key actions and the mood of a scene, they do not acknowledge repeated appearances of a character or object and, above all, they miss the intended meaning of our micro-narratives. As the application of automated image or video captions is relatively new territory to both human information retrieval and to human understanding in the context of media access, it is important to trace these observable phenomena back to source (their underlying problems). It is these issues which make current video captions appear trivial or naïve and which allow us to explore how human descriptive knowledge can potentially be applied to improve outcomes. We have therefore grouped the observed problems into three principal categories, each of which impacts the quality of outputs: methodological issues, where the problem is rooted in the nature of the training data; computer vision problems, which result from current limitations in object detection/identification; and linguistic problems, which are related to how the output of computer vision algorithms is rendered into natural language. Each area will be discussed below.

5.1 Methodological Issues

A significant problem is the nature of the available training datasets. In the field of image recognition and description a number of large, comparatively high quality, annotated datasets are available when compared to other types of training data (e.g. in the business world). However, these captioned image datasets are not optimised in a way that serves linguistic studies. This can be illustrated with reference to one of the principal training datasets used to create the first iteration descriptions for our MVC corpus, MS COCO (Lin *et al.*, 2015). As explained above, MS COCO is a meticulously designed and annotated large-scale dataset for visual object detection and captioning. Each still picture has been annotated with five captions, generated by five individual human operatives, describing the image content (Chen *et al.*, 2015). The purpose of this exercise is to harvest visually pertinent information from which machines can learn the connections between the visual objects and actions, and the semantic labels given to them by the annotators. As with other data-related tasks of a similar scale, the MS COCO creators resorted to crowdsourcing service Amazon Mechanical Turk to collect the image captions (Chen *et al.*, 2015). Although a widely accepted practice for manipulating datasets of this size, crowdsourcing annotations for training data in this manner introduces a number of factors which render the results from test data – in this case, our MVC corpus – less reliable, and demonstrably low in quality.

Firstly, the **type of work** undertaken is financially rewarded according to the number of units of material captioned, meaning that captions are produced spontaneously and rapidly, possibly without much thought being given to lexical variety or non-superficial observations.

The protocols attaching to such image captioning tasks include word count and time limitations, which can have a significant impact on creativity, resulting in rigid syntax.

Secondly, in terms of **workers and their profiles**, Amazon Mechanical Turk and similar crowdsourcing services tend to attract college students from a computing background, leading to age and interest bias (Difallah *et al.*, 2018). Research shows that the workers' profile has an impact on the quality of their work (Kazai *et al.* 2012) and that feedback can improve quality (Han *et al.* 2019). However, Chen *et al.* (2015) do not discuss the details of their approach to recruiting and working with the crowd workers, and the MS COCO captions suggest that at least some of the crowd workers are amateurs when it comes to the descriptive genre. The examples in Figure 11 illustrate the different skill levels. For instance, whilst caption 1.iii. sounds professional and forms a grammatically complete sentence with a verb in simple present, it includes an abstract value judgement ("beautiful"). Caption 1.iv. is factual but vague, giving little detail about the objects in the room ("lots of furniture"). Similarly, in image 2, several captions refer to the red sign, but lack the precise terminology (i.e. "no-entry sign") that may be needed in the context of content description for archival purposes or AD.

	1 (#374628) <ul style="list-style-type: none"> i. a kitchen made of mostly wood with a small desk with a laptop. ii. a full view of an open kitchen and dining area. iii. a beautiful, open kitchen and dining room area features an island in the center and wood cabinets and large windows. iv. a kitchen with wood floors and lots of furniture. v. a very spacious room with a kitchen and dining area.
	2 (#132394) <ul style="list-style-type: none"> i. a red sign is on the gray sidewalk ii. a vandalized street sign on a side walk iii. a red cautionary sign with "know hope" in graffiti iv. a round red sign on the other side of a stop sign v. a red sign is at the corner of the street on the sidewalk
	3 (#290868) <ul style="list-style-type: none"> i. a grandmother standing next to a child in a kitchen. ii. baby trying to open wooden cabinets under the sink. iii. a woman and child stand in the kitchen. iv. an older woman is standing in the kitchen with a child. v. the little girl is trying hard to open the cabinets

Figure 11: Examples of captioned images from MS COCO

The description **task** may also impact the quality of the results. The crowd workers for MS COCO were instructed to describe all "important parts" of the scene, using at least eight words, and not starting sentences with there is/are. An obvious problem is that crowd workers do not always follow the instructions. Albeit infrequently, they do use "there is/are"

(N=12817, see e.g. *Figure 5* above) and/or phrases such as “an image of”, “a full view of”, which are similarly redundant in this context. More importantly, the instruction rubric raises the highly relevant question: what are the “important parts” of any given image? Naturally, the answer is inextricably linked to matters of **relevance and saliency**. Considering image 1 in *Figure 4* again, each caption highlights different objects, illustrating the differences in human perception and approach to simple tasks of this kind. In a video scene, whether it is important to mention the laptop or to highlight the mostly wooden outlay will depend on the context of the unfolding narrative.

Further issues inherent in this type of description are **accuracy, vagueness and lexical ambiguity**. Chen *et al.* (2015) explore recall (i.e. whether an entity that is present in an image is referred to in the caption) and accuracy (i.e. whether the description is correct) for selected nouns, adjectives and verbs. Their results indicate high recall and accuracy rates for nouns denoting somewhat rare entities without many or any synonyms (e.g. “elephant”), but mixed rates for other more prosaic objects (e.g. “sidewalk”).

A more fundamental problem in our context is that although the aim of MS COCO was to present scenes, i.e. objects in context, it is still a database of **static images** without narrative coherence from one image to the next. As such, it can capture actions only to a limited extent and cannot provide examples of narrative cohesion (e.g. causal, temporal cohesion, links between characters, co-reference). As for actions, we clearly have ability to identify visual actions in still images, especially in photos, using common knowledge of body movements, postures etc. Thus MS COCO has numerous instances of walking, playing, drinking, which can be detected from a single frame. In addition, it contains verbs denoting actions that would stretch over several frames in a video scene, e.g. opening (Ronchi & Perona 2015), although these are considerably less frequent and occur in phrases such as “is trying to open”, suggesting uncertainty (see *Figure 11*, 3.ii and 3.v). Similarly, descriptions such as “he looks like he is falling”, although infrequent, indicate uncertainty in relation to such actions.

With regard to cohesion, **linkage of characters through actions** is limited and builds on a smaller set of verbs, mainly “talking”, but the frequent use of “talking” in our MD corpus is in itself problematic. It illustrates the point that human descriptions are narratively salient and relevant in a way that computer descriptions are generally not, at least consistently. When we see a man and a woman arguing about who does the washing up, narrative saliency may not to be found in the most common of computer captions, “A man and a woman are talking”. Adding a layer of emotional description may be possible if the computer determines facial expressions and therefore selects “A man and a woman are sad”, which might in a way indicate incompatibilities within the relationship. Most people would be able to detect this nuance by interpreting the dialogue in terms of the social setting, vocal tonality, facial expressions and body language. Meanwhile, the computer simply ‘sees’ two people talking. The computer may even reach this conclusion when the characters are not visibly speaking (i.e. their mouths do

not appear to be forming words). As a measure of quality, the value for the viewer is to be found in the storytelling and not in the quasi-metadata description represented as a formulaic ‘*man+woman+talk*’. In the example extract below, the salient point, for instance in parts 4 and 5, is not simply that ‘a man’ is talking but *what* he says and then the unimpressed look on the woman’s face is of importance. In the same vein, the point that the man finds the Ringo Starr album is of more importance than the fact that he is talking while browsing the records.

	Dialogue	Audio Description	Content Description	Machine Description (MD)	Images used in MD
1		Later they are in a music store.	Tom and Summer are in a music store, browsing. Summer looks unhappy.	A man is sitting in a library with a book shelf	
2	Tom: It pains me we live in a world where nobody has heard of <i>Spearmint</i> .			A woman is sitting on a couch and smiling	
3	Summer: I’ve never heard of them.		Summer sounds annoyed.	A man is sitting in a library and smiling	
4	Tom: I put them on the mixtape I made you. They’re track one.		Tom is surprised.	A woman is smiling and then she smiles	
5	Summer: Oh, yeah.		Summer looks uninterested.	A man is talking to a woman and she is smiling	
6		Summer nods, unconvincingly.	Tom rolls his eyes; he looks disappointed.	A man is talking to a woman and she is smiling	
7		Tom finds the <i>Ringo Starr</i> record.	Tom picks up a <i>Ringo Starr</i> record; he laughs and shows it to Summer. She smiles unimpressed.	A man is talking to a woman and she is smiling	
8		Summer gives a tight smile and walks away from the record stand. Tom reaches out to take her hand, but she pulls away.	Summer walks away; Tom follows her and tries to hold her hand. Summer moves away and Tom looks sad.	A man is talking to a woman in front of a bookshelf	
9		With a disappointed sigh, Tom follows Summer out of the shop.		A man is dancing in a room with other people	
10	[SFX WIND CHIME]		Summer opens the door and leaves the shop; Tom is right behind her.	A man is talking to a woman and she is smiling	
11	[SFX DOOR CLOSES]			A man is sitting in a chair and talking	

Figure 12: Clip #200115 with machine descriptions

Interestingly, while AD may assist in determining that a man and a woman are in the music store (the fact that they are not happy would be discernible to the viewer from voice tone and language), human content descriptions (CD) indicate everything that can be observed in the scene – two people, music store, music records, unimpressed faces, disappointment- falling short only on broader narrative interpretation, which requires material from outside that specific scene (the failing relationship, perhaps). To this extent, and for this particular purpose, the CD corpus can be considered a more appropriate and quality-driven resource.

One of the problems exacerbating the issue that there is no cohesion between individual MDs is also that the MD currently only describes the middle frame of each shot; the middle frame is not necessarily the most representative frame of a shot. This makes it even more difficult to create a coherent narrative.

The lack of linkage of characters is one indicator of the dataset’s limitations with regard to creating a cohesive narrative. Another indicator is the lack of **temporal, causal or other links between individual actions**, i.e. the absence of relevant cohesive markers. While ‘and then’ occurs within the MVC corpus, instances can be traced back to split-screen images in the training data which prompted captioners to treat them in sequence, belying the superficially temporal implications of the phraseology. Finally, narrative coherence is constructed in the way human beings identify, recognise and refer to characters. MS COCO, however, does not include any support for this, for example, in the form of cohesive chains drawing on pronominalisation and other ways to create **co-reference**. The absence of co-reference markers is certainly one of the most noticeable features in the current MD corpus. Many examples in which a series of captions refer to the same characters read as shown in *Figure 12* above. The story arc from which it is taken shows one man and one woman.

```
00:00:00.000 00:00:02.700 A man is talking and smiling and laughing
00:00:02.700 00:00:04.533 A woman is smiling and talking to someone
00:00:04.533 00:00:24.600 A man is dancing in a room with other people
00:00:24.600 00:00:26.733 A woman is sitting on a couch and smiling
00:00:26.733 00:00:28.266 A man is dancing in a room with a lot of people
00:00:28.267 00:00:30.734 A man is walking through a door and then he falls down
00:00:30.733 00:00:33.000 A woman is sitting on a couch and eating a sandwich
00:00:33.000 00:00:34.600 A man is talking and smiling and laughing
00:00:34.600 00:00:36.200 A man is sitting on a couch and talking
00:00:36.200 00:00:40.967 A man is talking and smiling and laughing
00:00:40.967 00:00:42.967 A woman is sitting on a bench and talking
```

Figure 13: Example of machine description from MVC clip #200006

Another difference is in the nature of the training dataset, i.e. a **mismatch between the content of the images in the training data and that of the MVC**. The images in MS COCO show simple everyday scenes of people walking, talking, eating, engaging in sports and so forth. The explicit aim of the MS COCO creators was to include non-iconic images, i.e. scenes without one person or object clearly standing out. In our corpus, which contains extracts from feature films, visual scenes are more deliberately composed, iconic and laden with narratively relevant *mise en scène*. They are also subject to editing techniques that manipulate visual content to include multiple shot changes, close-ups, panning and zooming techniques which render the material difficult for the machine to ‘read’.

Aside from the methods applied in relation to the purchase of training data captioning services from crowdsourced websites, and the differences in the nature of the visual material included in the training data and our MD corpus, other measures were taken during the application of the training data to MD production which impacted results. In particular, the lexical poverty of outputs was increased by the elimination of tokens in the training data which occurred fewer than four times. These ‘long tail’ words, being those which are uncommonly found in the corpus, are a regular feature of AD and human description adding nuance and colour. In this case, elimination from the training data before applying the *DeepCaption* model was a matter of computer processing expediency. Furthermore, topical bias is inherent in the types of data typically collected from *Flickr* and *Tumblr*, such that words like *laptop*, *microphone* and *surfboard* are over-represented in the test data results. Poor data cleansing within the training data also resulted in grammatical mistakes, lexical errors, and incomplete captions transferring across to the MVC machine descriptions. Finally, natural language processing as it has been applied to MD output, falls short of human descriptive requirements, being highly formulaic and syntactically repetitious in nature (“An X and a Y are *+verb gerund*”, as illustrated in the earlier examples). Taken together, these factors currently result in poor quality captions.

5.2 Computer Vision Problems

At the most fundamental level, visual storytelling relies on the successful identification of characters in order for the viewer to locate them successfully and consistently within the unfolding narrative. This is particularly the case for sight- and cognitively-impaired viewers, but also in the video retrieval scenario, where a certain character must be isolated from a vast wealth of video material. Separation between male and female protagonists where they are seen and not heard is generally helpful, notwithstanding issues of gender labelling and gender bias which are outside the scope of this study. Fully sighted human beings are capable of distinguishing between sexes featured in moving imagery in a traditional, binary sense with relative ease. The MD outputs from our computer model were unreliable in this regard, although the training data from which they were derived is unlikely to have had a significant error rate.

Still Image	Machine caption (MD)
	<p>Clip#: frame#: 200212:1560</p> <p>'A man is talking and smiling at someone'</p>

Figure 14: Example of incorrect gender assignment

In addition to the incorrect labelling of gender, in certain circumstances the inconclusive nature of the computer vision model leads to use of the phrase 'a person' to denote uncertain gender. This total number of 'a person' instances in the MD corpus is 139 (1976.82/m) whereas in training data this is a less frequent phenomenon (MS COCO: 3312.99/m; TGIF: 3240.83/m). A random sample of fifty concordances were examined to determine whether a pattern emerges. In forty-three of the concordances (86%) a part of someone's body was visible in the still image captioned (as opposed to the face, head or full body). Many of these examples contained hands holding something, or fingers.

Still Image	Caption details
	<p>MD corpus clip#:frame# 200508:93</p> <p>"A person is holding a cell phone in their hand"</p>

Figure 15: Example of 'a person'

AD containing incorrect labelling of male and female characters would be unhelpful at best, and at worst represent a significant confound for audiences experiencing sight-impairment. Vocal gender profiling work will undoubtedly help to rectify this issue, compensating for unreliable computer vision feature extraction which is currently too rigid and rule-bound (e.g. a person with short hair is generally labelled as a man, irrespective of dress, mannerisms, voice and other cues implying gender).

As an alternative approach, we have been liaising with consortium colleagues to test their vocal gender identification model on our feature film material. This builds on the work undertaken within the project to diarise multimodal voice outputs as a preliminary step to calculating male/female gender split in the specific case of the French audiovisual landscape (Doukhan *et al.*, 2018). Vocalisation techniques – for example, patterns in the expulsion of air during speech, and gender-specific pitch of vowel formants – are used to profile vocal tracks and determine the sex of the speaker (Doukhan *et al.*, 2018). However, this process is not optimised for the English language. Nevertheless, the machine adapted to natural English to a limited extent, while extremes of emotion (e.g. crying, shouting) created a confound. The feature film genre also presented problems for the model, which was trained and evaluated on news, interviews and debates, such that extraneous noise (e.g. street sounds, music mixed with speech) reduced the efficacy of speech analysis. We expect to undertake further research in this area during the life of the project, and intend to report the results of this work in the next deliverable (D5.3).

Similarly, machine-based object detection remains unreliable to the extent that non-standard angles, changes of size/scale and rapid changes of light and shade can alter the description from ‘a car’ to ‘a guitar’ between one frame and the next, or can elicit an object description in one frame but not in a subsequent frame. For instance, example X (= example used in 3.5.1 to show what MD looks like), included one instance of the word ‘kite’ (although, as another computer vision problem, the object denoted as a kite is in fact a shield shaped sign). In the image where the word ‘kite’ is used, the sign is seen frontally, in the other image, it has a more unusual angle and the caption makes no reference to it.



Figure 16: Example of ‘Kite’ in MD corpus

Search for ‘kite’ in the MS COCO dataset reveals the variety of images showing a kite (six examples shown below).







	<p>221291 http://farm1.staticflickr.com/254/519398801_f9b8e32a24_z.jpg a little boy standing in the grass with a kite in the sky in the background. a little boy standing in a field below a kite. a young boy is posing in a large grassy area. a boy is out on the park flying a kite young boy posing in front of a flying kite in the park</p>
	<p>154520 http://farm3.staticflickr.com/2753/4434449872_4f2ca42f20_z.jpg there is a man holding on the a kite that hes flying a large kite is flying in the sky a man flies a kite on a sunny day a beautiful clear blue sky is ideal for flying his kite. a person is under a clear sky flying a rainbow kite.</p>
	<p>132328 http://farm5.staticflickr.com/4070/4468423978_f2ff27701e_z.jpg a man and his son fly kits in a field as a crowd watches. a group of people playing with kites in the park on a sunny day many people watch a person fly a kite with a young person a father helps his son fly his kite. the father and son are looking at the kites flying overhead.</p>
	<p>348982 http://farm4.staticflickr.com/3561/3393764736_8baeb962aa_z.jpg the lady holds a small box kite on a string. a lady doing something interesting with some kite in cold weather. a woman in a brown jacket holding a kite in a field. a woman holds bags and a kite that resembles two boxes. a woman is holding a kite in a park.</p>
	<p>041859 http://farm3.staticflickr.com/2371/2084946944_9e4c065868_z.jpg illustration of a silhouetted person with a kite. a painting of a person walking with a trailing kite. painting of a child with a kite in an orange sky a painting of a person walking along a field holding a kite. a child is flying a kite in this drawing</p>
	<p>160239 http://farm5.staticflickr.com/4083/5029487385_08bc30de4b_z.jpg very large balloon depicting whale on display at beach. a kite fashioned to look like a whale on a beach. a large inflatable whale sitting on top of a beach. this is a whale balloon in a parking lot a large whale kite some buildings and people</p>

Figure 17: Example of 'Kite' in MS COCO

In a similar vein, the machine is not currently able to extract facial expressions from multimodal material, i.e. laughing and smiling cannot be detected or distinguished from each other in the current model:


Still Image	Caption
	MD Clip:image# 001603:833 'A woman is smiling and laughing while wearing a black dress.'

Figure 18: Example of 'smiling and laughing' (incorrect)

The overall number of the token 'smiling' in the MD corpus is 1,654 (23,522.72 per million), 3,096 in MS COCO (445.22 per million), and 5,522 in TGIF (4,147.36 per million). In a random sample of fifty concordances containing 'smiling', twenty-seven (54%) of the identified characters are not smiling, but rather frowning or grimacing. In most cases, the common denominator is the presence of at least one face in 'close up':


Still Image	Caption
	MD Clip:image# 200810:306 'A man in a suit is smiling and talking'.

Figure 19: Example of 'smiling and talking' (incorrect)

The phrase 'smiling and laughing' appears 68 times in the MD corpus (967.08 per million), but only 4 times in the MS COCO dataset (0.58 per million), and 194 times in TGIF (145.71 per million). Clearly this is a significant over-representation and is likely to represent some aspect of over-compensation in the features extraction, which might be investigated by the Aalto team. In a randomised sample of 50 'smiling and laughing' concordances the machine was mistaken on 32 occasions (i.e. 64%).



Still Image	Caption
	MD Clip:image# 003304:1419 'A man is smiling and laughing at something'.
	MD Clip:image# 103810:1173 'A woman is smiling and laughing while she is talking'.

Figure 20: Example of 'smiling and laughing' (incorrect)

While feature extraction and more training data is required to overcome some of these facial recognition difficulties, again, audio cues could possibly assist if incorporated into the model, as noted above.

In a similar vein, 18 instances of 'serious look' can be found in the MD corpus (255.99 per million), 15 in MS COCO (2.16 per million), and 48 in TGIF (36.05 per million). Although over-represented as a proportion of the MD corpus – suggesting perhaps that this was not the most narratively salient feature in the frame, but simply the one that the computer was best trained to extract – almost all instances of 'serious look' were correct:


Still Image	Caption
	MD Clip:image# 100705:1136 'A man is walking in a room with a serious look on his face'.

Figure 21: Example of 'serious look'

The word 'surprise' is used twice in the MD corpus, with both taking the form 'A man is walking through a door and is surprised by a woman' (#201614 and #204010). In the first example

(#201614), *Figure 22 (i)* below, the man's face is not visible in the captioned frame, pointing to the conclusion that the element of surprise is not detected via facial expression. The second example, (#204010), *Figure 22 (ii)* below, contains an image of two men standing in front of a window, with no suggesting of surprise on their faces, or indeed, the presence of a woman. Once again, there seems no immediate correspondence between visually expressed emotion and the machine-generated caption.



Still Image	Caption
	<p>(i)</p> <p>MD Clip:image# 201614:308</p> <p>'A man is walking through a door and is surprised by a woman'.</p>
	<p>(ii)</p> <p>MD Clip:image# 204010:308</p> <p>'A man is walking through a door and is surprised by a woman'.</p>

Figure 22: Examples of 'surprised'

Finally, the concept of 'making faces' is present in the MD corpus but can only be found a total of six times (85.33 per million). While this is not a single facial expression *per se*, use of the phrase implies some visual facial recognition acuity in the machine outputs. It is not possible to determine why, for example, the couple's faces in the first image (below) which appear to warrant the caption 'smiling' or 'laughing', are captioned '*a man ... is making faces*'. While the answer undoubtedly lies in the training data, since both options are available, it might be expected that poor feature extraction is in fact the source of the problem in this instance.





Still Image	Caption
	
MD Clip:image# 100305:2123	MD Clip:image# 101206:269
	
MD Clip:image# 101206:507	MD Clip:image# 202813:603
All frames captioned: 'A man is sitting on a couch and making faces.'	

Figure 23: Examples of 'A man is sitting on a couch and making faces'

In computer vision terms, facial expression detection is closely related to the rendering of emotion in film more generally. Obviously, a situation or scene might be regarded as 'happy' even though protagonists' faces do not exemplify the fact. It is remarkable that of the seven basic human emotions (Ekman & Friesen, 2003), only 'surprised' is present in the MD corpus, especially as the TGIF corpus (and to a lesser extent MS COCO) contains all except 'contempt'. All seven emotions were present in both the MS COCO and TGIF datasets, with the following frequencies recorded:

Emotion	MS COCO	TGIF	MD corpus
happy	63.56	214.05	0
sad	16.25	232.08	0
angry	11.5	168.24	0
disgusted	0.43	31.54	0
afraid	0.86	12.02	0
surprised	7.33	72.1	28.4
contempt(uous)	0	0	0

Figure 24: Relative frequencies of basic emotions per million

A significant problem with training the computer to determine emotional temperature in film is the requirement for close-up facial shots on the one hand, and audio markers (happy music, the sound of crying) on the other. Even where one or more of these is present at the narratively salient juncture, training the model to analyse facial features and incorporate sound cues simultaneously is far outside current reach.

5.3 Linguistic Considerations

As discussed above, the source of training data captions has resulted in MD lexical poverty in both variety and nuance. A study of verb usage in the MD corpus serves to illustrate this point:

<i>MD Corpus Verb Rank</i>	<i>Lemma</i>	<i>Frequency</i>	<i>MD Corpus Verb Rank</i>	<i>Lemma</i>	<i>Frequency</i>
1	be	7806	24	live	51
2	talk	1686	25	wear	48
3	smile	1682	26	smoke	46
4	look	1657	27	run	42
5	dance	1119	28	make	38
6	walk	1087	29	eat	24
7	sit	1004	30	pour	20
8	kiss	328	31	blow	16
9	hold	302	32	take	15
10	play	238	33	swim	14
11	drive	230	34	do	14
12	fall	214	35	fly	13
13	stop	203	36	work	13
14	sing	179	37	wave	13
15	stand	134	38	move	13
16	jump	130	39	read	11
17	laugh	79	40	open	10
18	put	73	41	hug	9
19	turn	72	42	cut	8
20	lay	61	43	show	8
21	lie	55	44	crash	5
22	ride	52	45	type	5
23	drink	51	46	park	5



Figure 25: MD Corpus: Verb Rankings

Eighty-eight verb lemmas can be found in the MD lexicon, only forty-six of which occur five or more times (see Figure 25). The most commonly used verb lemma is 'be' (frequency: 7806; relative frequency: 111.014.72/million), in contrast with the British National Corpus, which shows a relative frequency of around one-third of this rate (36762.66/m). In the MD corpus, 7549 instances of this lemma register in the third person singular (96.7%). Furthermore, 7508 of the 7549 instances of 'is' in the MD corpus are to be found in concordance with a corresponding verb gerund (CQL search: [word="is"&word=".ing"]), e.g. "A woman is dancing", "A man is talking", and so forth. Parsing during the NLP phase of image processing might be improved to provide more syntactic variety in the rendering of these machine descriptions.

In addition, the top six verb lemmata are vastly over-represented in the MD outputs when compared to the MS COCO and TGIF training datasets (Figure 26), suggesting that feature extraction and other factors play a significant role.

RANK	VERB LEMMA	MD f	MD verb/m	COCO f	COCO verb/m	TGIF f	TGIF verb/m
1	Be	7806	111014.72	154295	22188.44	90737	68149.11
2	Talk	1686	23977.81	3114	447.81	5914	4441.78
3	Smile	1682	23920.93	3913	562.71	3755	2820.24
4	Look	1657	23565.38	16902	2430.6	11071	8315.01
5	Dance	1119	15914.1	67	9.63	2392	1796.54
6	Walk	1087	15459.01	17921	2577.14	6480	4866.88
7	Sit	1004	14278.6	68705	9880.15	5076	3812.39
8	Kiss	328	4664.72	165	23.73	3242	2434.94
9	Hold	302	4294.96	30487	4384.19	5613	4215.71
10	Play	238	3384.77	15935	2291.54	4469	3356.5

Figure 26: MD Verbs: Comparative Statistics vs. Training Datasets

	<p>003693 http://farm2.staticflickr.com/1235/856828929_8055bb6a26_z.jpg people dancing and hanging out talking looking at their phones. a group of teenagers standing by a graffiti'd wall. some people and the male is wearing a gray shirt several teens in a concrete area one looks as though he is preparing to dance a young man looking at his feet with four pretty women in the background.</p>
	<p>557564 http://farm7.staticflickr.com/6102/6221836674_6822d45dc8_z.jpg the man in the picture is getting ready to dance. a man in a suit and tie wearing a hat. black and white photograph of a man in a business suit and hat a man wearing a suit and tie with a hat on his head. a man dressed in business attire and wearing a fedora.</p>

	<p>470467 http://farm8.staticflickr.com/7208/6957209805_241256bdd9_z.jpg two young women perform a dance in elaborate dress. two asian women doing a dance and one holding an umbrella. two oriental women appearing to dance, one with a big umbrella. japanese dancers, in costume, performing on a stage. two women perform a traditional dance on stage.</p>
	<p>438294 http://farm1.staticflickr.com/166/341488481_3f53299ed9_z.jpg a woman with her arms up while playing a video game a woman holding her arms in the air while holding a wii controller. a couple of women play a video game woman standing in front of chair holding a game controller. a woman in a black paisley skirt is trying to dance.</p>
	<p>169172 http://farm8.staticflickr.com/7221/7328349656_cd94ba6bbe_z.jpg people sitting down watching a couple dance in front dancers performing in front of an audience in a house. a bunch of people that are in a living room. two people dance in a room at night while an audience sits and watches. a man and a women are entertaining people by demonstrating either martial arts or dance.</p>
	<p>335587 http://farm4.staticflickr.com/3401/3207795270_cccfae1d67_z.jpg a male in a tie and black shirt and a white wall a man wearing a black shirt is dancing in front of an object. a man in a tie smiles and pumps his fist. a man standing in front of pile of reflective ribbons. a man that is wearing a shiny tie and dancing.</p>
	<p>116149 http://farm8.staticflickr.com/7009/6741036563_c7b4d25392_z.jpg a group of women dancing with umbrellas in a play. a group of four different women with umbrellas. oriental dancers dressed in blue holding blue umbrellas. some woman standing on stage doing a dance with umbrellas a group of four umbrella twirlers performing in a show.</p>
	<p>073665 http://farm4.staticflickr.com/3209/2876022288_cb57e117c7_z.jpg a man practices skateboarding in front of a building and parked cars. a young man does a skateboard trick on a city street. man on street appearing to be dancing or skipping sideways. . a boy on a skateboard in the air above a street a man flips his skate board on a city street.</p>

	<p>506874 http://farm5.staticflickr.com/4127/4988570358_59db919c46_z.jpg people sit under a red umbrella to watch pow wow dancers. a group of people at a party sitting next to a red umbrella. the woman is holding an umbrella at a festival. people sitting by a red umbrella take in an outdoor show. someone wearing a full body headdress is dancing around in front of a crowd.</p>
	<p>361265 http://farm8.staticflickr.com/7230/7287952600_534edd9135_z.jpg a woman dances across a brick street while holding an umbrella. a woman crossing a street holding an umbrella a woman walking across a street holding a camera. a woman posing on the street for a photo the ladies dancing happily in the street with an umbrella.</p>

Figure 27 (a): Examples of ‘dance/dancing’ from MS COCO



<s value="t0 000001:50">A woman is dancing in a room with a lot of people</s>



<s value="t1 000001:188">A television is showing a man on a television</s>



<s value="t2 000001:317">A man is dancing in a room with a lot of people</s>



<s value="t3 000001:393">A man is dancing in a room with a lot of people</s>



<s value="t4 000001:442">A woman is holding a box and talking to a man</s>



<s value="t5 000001:490">A woman is holding a large colorful kite</s>



<s value="t6 000001:546">A woman is holding a box and dancing</s>



<s value="t7 000001:590">A man is playing a piano and singing</s>



<s value="t8 000001:643">A man is dancing in front of a group of people</s>

Figure 27 (b): Example of ‘dancing’ from the MD corpus

In this example we can observe a randomised application of the gerund ‘*dancing*’ within the machine captions (see 3.5.1., above), which in turn is both correct and incorrect, absent and present:

- t0 – ‘dancing’ is used correctly; ‘with a lot of people’ is incorrect
- t1 – dancing is present on the television screen, but not in the caption
- t2 – incorrect; arms are outstretched but the character is carrying a parcel, not dancing
- t3 – correct/incorrect; character is possibly dancing, but is not a ‘man’
- t4 & 5 – no dancing visible in the image or present in the caption
- t6 – no dancing is present in image, but is present in the caption
- t7 – no obvious dancing in image (other than possibly on the tv) and no ‘dancing’ in captions
- t8 – incorrect; the girl on the left of the image is standing with arms raised, and on first inspection does not appear to be dancing, but walking towards the present

Clearly, the computer model generates captioning that can be both correct and incorrect in relatively similar and proximally close (sequential) visual circumstances. There would appear to be a preponderance of arms visible in captions containing the ‘dancing’ verb, however this is by no means a failsafe rule. Again, pixel level similarities between images occurring in the training data and those in the test data (MVC) are likely to be the principle causal factor for such anomalies.

An alternative source of information about the skewed nature of MD outputs are keywords. They provide score-based data regarding the uniqueness of the focus corpus in relation to a more generic and linguistically typical reference corpus. For this purpose, our comparison was made between the MD lexicon and that of the British National Corpus (BNC) which contains in excess of 96 million words, 6 million sentences, 1.5 million paragraphs and 700,000 unique items.

Analysis of keyness within the MD corpus illustrates the nature of lexical bias found within the captioned training data. Keyness is denoted by ‘keywords’, which have been described as: “words (single-token items) that appear more frequently in the focus corpus than in the reference corpus. They can be used to identify what is specific to one corpus (focus corpus) ... in comparison with another corpus (reference corpus)” (Sketch Engine, undated). In particular, the sources of imagery in the adopted datasets, which were derived from Flickr (in the case of MS COCO) and social media postings (TGIF), led to a preponderance of objects which were over-represented when compared with the more standard lexicon in the reference corpus (BNC). Technology and youth-relevant vocabulary scores highly in MD keyness with *laptop*, *skateboard*, *trampoline* all ranking in ‘top 5’ positions; *tv*, *microphone* and *piano* fall within the ‘top 20’ items; and *surfboard*, *motorcycle*, *guitar*, and *skateboarding* rank in the ‘top 30’. These scores illustrate the youth and technology bias generally observed within social media postings and thus are over-represented in the training data. The over-represented nature of *hallway* (rank:1; frequency 305; relative frequency: 4337.62/m) appears to derive from a particular phenomenon in the training data. Of the 305 occurrences in the MD corpus, 255 can be found in the concordance ‘walking down a hallway’, suggesting similar concordances

occur in the training data. Indeed, while this phrase appears only five times in the COCO dataset, it can be found 65 times in the TGIF dataset (48.82/m).

Clearly, the disparity in relative frequencies between the MD corpus and training data suggests that a level of bias is being introduced via the *DeepCaption* model, which requires further investigation. *Couch*, as the second ranked item in order of keyness, occurs 306 times in the MD corpus, with a relative frequency of 4351.85/m. A total of 296 of these MD occurrences feature in the concordance ‘sitting on a couch’ (relative frequency: 4223.85/m) and ‘sitting on a couch and smiling’ occurs 82 times (relative frequency: 1166.18/m). In the COCO dataset, ‘sitting on a couch’ appears 872 times (relative frequency: 125.4/m), whereas in the TGIF dataset, it can be found 217 times (relative frequency: 162.98/m). Again, the imbalance between training data and MD corpus suggests that commonly occurring phrases become over-represented during the captioning process.

Rank	Term	Score	(MD) corpus frequency	Reference corpus (BNC) frequency
1	hallway	920.81	305	417
2	couch	596.12	306	708
3	laptop	458.46	60	97
4	skateboard	355	42	77
5	trampoline	321.45	34	57
6	dance	286.34	1119	6132
7	smile	154.75	1682	17255
8	tv	118.73	14	77
9	singing	108.56	91	1228
10	shirtless	106.26	8	9

Figure 28: MD corpus, keyness scores

As a further illustration of lexical poverty, only 17* adjective-tokens are present in the MD corpus, compared with 568 tokens in the TGIF and 1566 tokens in the MS COCO datasets. Although longtail words have been removed for the purpose of image caption processing expediency, a mere 39 adjective-tokens are found in the MD corpus when the minimum frequency is reset to ‘1’. Thus, there would seem to be a significant disconnect between the training data and focus corpus in this regard, something which warrants further investigation in terms of feature extraction at the level of model building.

adjective	MD (f)	MD/m	COCO/m	TGIF/m
dark	108	1535.95	212.98	1635.81
other	92	1308.4	1828.48	3596.84
white	36	511.98	4692.51	3799.62
long	35	497.76	468.66	2111.99
large	21	298.66	3669.48	690.23
serious	18	255.99	9.49	158.47
black	17	241.77	2642.42	4344.89
next	15	213.33	5577.35	768.34
young	15	213.33	2536.87	7633.79
red	12	170.66	2225.1	1897.93
green	9	128	1351.19	467.16
shirtless	8	113.77	51.34	210.3
laptop	8	113.77	537.11	10.51
video	7	99.55	253.53	0
remote	6	85.33	356.21	23.28
wooden	6	85.33	1346.45	108.9
smart	5	71.11	48.61	16.52

Figure 29: Adjectives: MS Corpus vs. MS COCO and TGIF Training Datasets* (*minimum frequency n=5)

Colour, as one sub-group of adjectives applied across the corpus, was studied in its own right. We wanted to discover if the machine was capable of determining colour from the supplied images, or whether colour was a function of training data collocations. Overall, 80 captions in the MD corpus contain colour collocations:

Colour	MD (f)/m	Microsoft COCO (f)/m	TGIF (f)/m	Correct (MD)	Incorrect (MD)
Yellow	3 (42.67/m)	8,749 (1,258.15/m)	643 (482.93/m)	2	1
Blue	3 (42.67/m)	14,352 (2,063.89/m)	1,776 (1,333.89/m)	1	2
Green	9 (128/m)	13,566 (1,950.86/m)	660 (495.7/m)	2	7
Red	12 (170.66/m)	18,114 (2,604.89/m)	2,664 (2,000.83/m)	6	6
Black	17 (241.77/m)	20,294 (2,918.39/m)	6,326 (4,751.22/m)	4	13
White	36 (511.98/m)	37,924 (5,453.67/m)	5,286 (3,970.11/m)	22	14
Total	80	-	-	37	43
%	-	-	-	(46%)	(54%)

Figure 30: Colour representations in MD, MS COCO and TGIF

Statistically, the results are inconclusive about the capabilities of the model in relation to colour recognition (46% correct, 54% incorrect). However, it is possible to determine that many of the colour choices relate to captions gleaned specifically from the MS COCO training data. To illustrate this point, there are 83 occurrences of the collocation ‘yellow surfboard’ and 78 examples of ‘blue surfboard’ in MS COCO data. Neither of these collocations appear in the TGIF dataset. Yet the MD corpus contains the following caption:

	<p>(i)</p> <p>MD Clip:image# 004102:50 ‘A yellow and blue surfboard sitting on top of a wooden table.’</p>
	<p>(ii)</p> <p>MD Clip:image# 200212:158 ‘A man standing next to a large white and green airplane.’</p>
	<p>(iii)</p> <p>MD Clip:image# 201311:1303 ‘A man is smiling and laughing while wearing a black shirt’ COCO x530 TGIF x772</p>

Figure 31: Colour representations

We can conclude, in this case, that the MD caption has been derived entirely from the MS COCO corpus, via the amalgamation of two phrases within the dataset.

The MD caption for image #200212:158 (above) can be directly traced back to MS COCO, which registers six instances of ‘green airplane’, while TGIF contains none. Clearly, the image contains neither an airplane nor obvious green colour or tonality. In this instance, we must look to other explanations for the caption selection, with feature extraction being a likely contender.

Finally, caption #201311:1303, ‘A man is smiling and laughing while wearing a black shirt’ illustrates an interesting point about colour captioning in relation to the training data. In this case, three examples of the phrase ‘black shirt’ occur in the MD corpus, while MS COCO contains 530 instances, and TGIF, 772. Yet the man in the image is wearing a red shirt. The natural conclusion would be that ‘wearing a red shirt’ does not occur in the training data and

therefore is not available to the machine to use in captioning. However, ‘wearing a red shirt’ has an occurrence of TGIF:59 and MS COCO:71. If the computer were able to identify colours, then it would be reasonable to expect the caption would draw from the colour specific information in the training data. Since this was not the case, and the computer selected ‘black shirt’, one might presume that other visual features were used in order to select an appropriate caption. In this case, linguistic analysis seems to confirm our earlier suspicions that the machine does not ‘see’ colour, using alternative parameters within the image upon which to reference the final description. The finding has consequences that reach beyond colour recognition purely as an image descriptor, because character identification and cohesion between frames and scenes in long-form narrative are often reliant on costume as a marker denoting continuity. If the man in *Figure 22* were to be seen from behind in the next frame, the human describer would infer it to be the same person, based on his clothes, body shape, and perhaps hair style – even though we may have seen none of these from a rear view previously. Recognising colour is therefore a key task in building a computer model that is capable of sequencing narrative, although once recognised the question of colour saliency will need to be addressed (e.g. does the colour of person’s shirt help to identify them as the protagonist at the centre of narrative in some circumstances; and is the colour of a particular object salient to the plot?). In other words, colour identification for sequencing purposes may be regarded as a different proposition from colour recognition for purely artistic, or narratively important purposes.

6 Video Sequencing

The viewer constructs coherence in storytelling from a wide range of cues, some more readily accessible than others. The human mind works hard to make sense of continuity clues in film narrative, with factors such as location (e.g. who lives in a particular house or works in an office interior), clothing (is the girl in the red coat on the train the same girl who is now climbing the stairs?), body silhouette and posture (we would be unlikely to confuse two protagonists, if one was shaped like a wrestler and the other a long-distance runner, even when they are filmed from a distance, or from an obscure angle). When a human recognises these markers, assumptions are made about the nature of the person at the centre of the narrative action, even where these assumptions are later discarded in light of subsequent information. Successive actions are read as a continuity of plot, and weighed up accordingly.⁴

In audio description, pronouns are used frequently as a form of shorthand (given the lack of sufficient hiatuses for long explanations) to make aspects of narrative action and the cohesion of character appearances between frames and scenes more readily accessible. Pronouns allow the human describer to avoid cumbersome and repetitious reference to characters by name or other referring expressions (e.g. complex noun phrases such as *the tall woman who just*

⁴ Theoretical explanations for the human capacity to process and makes sense of textual and multimodal input were discussed in more detail in Deliverable 5.1.

entered the room), multiple times in close succession. By doing so, the cognitive load on the viewer is greatly reduced.

Currently, the machine does not have the tools, or indeed visual vocabulary, to assign character continuity markers to protagonists in moving images, not least because machines still deal in the currency of single, still frames. While the computer certainly cannot connect individual frames yet, the fact that machine descriptions draw upon human-crafted training data produces some pronominalisation within the MD captions (as opposed to linking individual captions).

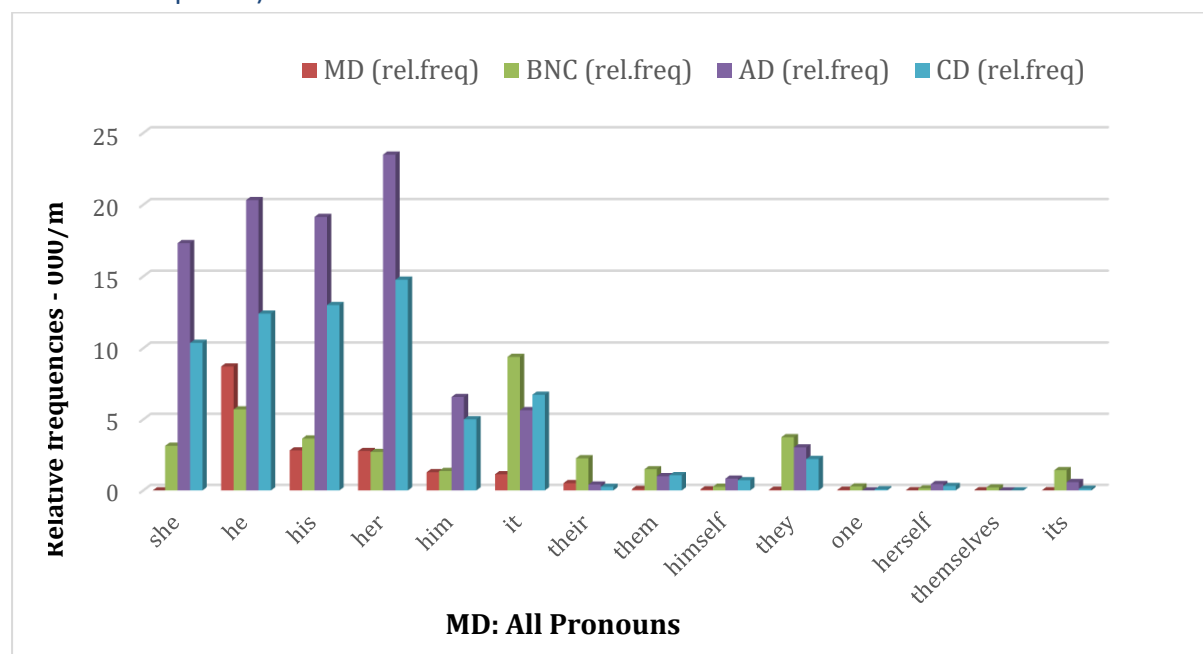


Figure 32: MD Pronouns: Relative Frequencies vs. BNC, AD and

*used as a pronoun only

Nonetheless, it is to be expected that pronoun usage in the MD corpus would fall below that in the BNC, given that personal pronouns such as 'I' and 'we', 'my' and 'ours' and so forth were excluded from the captions during training data compilation. 'He' and 'she', 'his' and 'her' were permitted in the training data, and represent valuable cohesive devices in the narrative context, but are vastly under-represented in the MD corpus (see graph in Figure 32).

Similarly, the distribution of articles across the different corpora indicates differences in referential identification as a further cohesive device. The indefinite article, which indicates that the referent at hand is treated as new, is strongly over-represented in the MD corpus, meaning that a large number of referents are introduced/treated as new. By contrast, the other corpora have a much stronger representation of articles that are used to mark referents as 'known' (A man enters the room. Then the (same man) does X.), 'inferable' (A car appears. The driver... [cars have drivers; part of a scenario we activate]) or 'situationally evoked' (The driver gets out and goes to the shop around the corner [salient in the context]).

Token	MD <i>f</i> (/m)	CD <i>f</i> (/m)	AD <i>f</i> (/m)	BNC <i>f</i> (/m)
a	14,704 (209,116.12)	1,777 (34,483.43)	948 (32,636.76)	2,163,730 (19,259.576)
the	263 (3,740.31)	2,898 (56,236.9)	1,739 (59,868.49)	6,054,950 (53,895.69)
this	0	7 (135.84)	4 (137.71)	454,536 (4,045.87)
that	6 (85.33)	44 (853.84)	16 (550.83)	1,120,808 (9,976.42)
these	0	0	0	123,624 (1,100.39)
those	0	0	0	87,197 (776.15)

Figure 33: Frequency and relative frequency of articles in different corpora

Among the tokens in the above table, only ‘a’ appeared as a keyword across all corpora when observing the high frequency keyness in *Sketch Engine*. Other tokens in the list do not appear of high frequency keywords in any corpora. Keyness for higher frequency words, in this case (see table below) is measured by scoring the tokens in MD against our sub-corpora, AD, CD and the benchmark BNC:

Reference Corpus	Score
CD	5.92
AD	6.25
BNC	10.37

Figure 34: Keyness score of the article ‘a’ across sub-corpora with MD as focus corpus

Part B

1 Introductory remarks

This part of D5.2 presents the results of a qualitative analysis of human content descriptions created by archive editors for the purposes of search and retrieval within the media archive of Finnish public broadcaster Yle. The aim of the analysis was to gain an understanding of the archive content description process: to find out (i) how visual information is described for archival purposes and (ii) which factors affect the description. In addition, the content descriptions produced by the archive editors are compared with audio description, with the aim of investigating whether automated, or semi-automated description production might serve these users. As this study was conducted with just one company, Yle, it is important to note that media houses may differ in the practices of describing the audiovisual content and the purposes of the description.

For this analysis, eighteen programmes representing different genres and programme types were selected from the Yle dataset. Each programme had been previously annotated with content descriptions (CDs) by a production coordinator or an archive editor. In addition, interviews with senior content describers were conducted and guidelines for the creation of CDs discussed.

Section 2 presents the data more in detail. Based on guidelines and interviews, section 3 provides background information on principles and practices in content description at Yle. In section 4, content description of one programme is analysed in detail, and section 5 presents the results of our analysis of all eighteen programmes. In the final section, 6, the main findings are summarised.

2 Data Analysis

Our plan for data analysis incorporated three key aspects:

- 1) Consideration of the Yle guidelines as a framework for content description, informing the types and granularity of content description according to programme type and re-use value
- 2) Interviews conducted with three professional archive editors and three production coordinators employed by Yle to discuss workflows and decision-making process when creating CDs
- 3) Metadata (including content descriptions and subtitles) corresponding with the video material retrieved from the corpus of eighteen television programmes (detailed in Figure 35, below) reviewed in parallel with moving imagery.

A table containing the full corpus inventory can be found in Figure 35, below.

Title	Classification Main / Sub	Programme	Duration	First run / Archive date	Yle_dataset
Yle Uutiset Kaakkois-Suomi 'Yle News South-East Finland'	News / magazine	MEDIA_2014_00778459	10M19S	2014-05-20 / 2014-07-30	004_may2014
Yle Uutiset Pohjanmaa 'Yle News Ostrobothnia'	News / magazine	MEDIA_2014_00780100	6M3S	2014-05-23 / 2014-06-13	004_may2014
Yle Uutiset Suora linja 'Yle News Direct line'	Current affairs / magazine	MEDIA_2014_00778852	8M38S	2014-05-21 / 2014-06-02	004_may2014
MOT 'QED'	Current affairs / report	MEDIA_2009_00019261	29M19S	2009-09-21 / 2010-01-04	11_EnglishSubs
Silminnäkijä 'Eyewitness'	Current affairs / report	MEDIA_2013_00679113	27M53S	2013-10-24 / 2013-12-17	11_EnglishSubs
Närbild 'Close-up'	Current affairs / magazine	MEDIA_2014_00775849	28M30S	2014-05-19 / 2014-06-30	004_may2014
Eurovaalit 2014 'EU-election 2014'	Current affairs / discussion, interview	MEDIA_2014_00778940	38M48S	2014-05-21 / 2014-06-09	004_may2014
Dokumenttiprojekti 'Document project'	Factual / document	MEDIA_2014_00720911	48M13S	2014-02-02 / 2014-06-26	11_EnglishSubs
Sohvasurffaajat 'Sofa surfers'	Factual / document	MEDIA_2013_00626728	27M46S	2013-07-02 / 2014-01-17	11_EnglishSubs
To Nightwish with Love	Factual / document	MEDIA_2016_01145263	58M17S	2016-08-20 / 2017-02-15	11_EnglishSubs
Ulkolinja 'International line'	Factual / document	MEDIA_2018_01418093	53M23S	2018-04-12 / 2018-06-28	11_EnglishSubs
Sissipuutarhurit 'Guerrilla gardeners'	Factual / report	MEDIA_2012_00460971	27M30S	2012-07-05 / 2012-08-08	11_EnglishSubs
Tekijänä 'Made by'	Factual / report	MEDIA_2012_00395438	28M28S	2012-03-04 / 2012-06-19	11_EnglishSubs
Puoli seitsemän 'Half past six'	Factual / discussion, interview	MEDIA_2014_00772163	28M24S	2014-05-07 / 2014-05-27	004_may2014

Strömsö	Factual / infotainment	MEDIA_2017_ 01221354	28M43S	2017-02-05 / 2017-05-24	001_Stromso01 –
Strömsö	Factual / infotainment	MEDIA_2017_ 01225114	28M38S	2017-02-12 / 2017-06-06	001_Stromso01
Strömsö	Factual / infotainment	MEDIA_2017_ 01235000	28M47S	2017-03-05 / 2017-06-05	001_Stromso01
Strömsö	Factual / infotainment	MEDIA_2017_ 01355102	28M07S	2017-12-03 / 2017-12-14	001_Stromso01

Figure 35: The Yle ‘Programme Corpus’ inventory

3 Content description at Yle

In this section, the guiding principles for content description at Yle are described on the basis of Yle guidelines and the interviews with production coordinators and archive editors.

‘Production coordinator’ is a traditional media role that sits within the team of creatives responsible for making a television programme, and therefore lies outside the archive team structure. However, Yle operates a system whereby production coordinators are responsible for content description, i.e. the descriptions are mostly created within and by the production team. The archive editors’ role is to check the descriptions (their overall structure) and give feedback and instructions to the production teams. The latter also describe programmes which pre-date current archival workflows having been in the archive for some considerable time. They follow a different pattern of annotation and documentation.

The purpose of content description at Yle is to enable re-use of the content in an effective and commercially focused way. Re-sale is dependent on a number of factors which ultimately guide the material contained in the description: the rights to reuse the content in other programmes, the usability of the extract (for example the quality of the image, the length of the shot) and the usability of the description for effective search from the archive. The description of segments has two purposes: to point out images with re-use potential (i.e. those which can be inserted in other programmes) and, on the other hand, to mark footage which is contractually restricted in terms of re-use or re-sale.

According to Yle guidelines, the *visual* content (what is visible in the image) is described in the content descriptions. In the search, other strata (time-coded metadata of the programme) such as key images, subjects, subtitles etc. can also be used. Accordingly, information which is described in the other strata should not be included in the description of the visual information. So, speech which is captioned in the subtitles should not be quoted, although the guidelines also say that its content can be briefly described, if it is regarded as important.

If a programme consists of several sub-subjects, the material will be segmented to facilitate retrieval. A segment is a thematic coherent whole; when the subject, location or target changes, the segment also changes. Furthermore, the visual content is segmented into parts every 5-10 minutes even if there are no changes.

As a general rule the visual analysis, which is rendered as content description, should answer the following questions:

- Who says or does something?
- What / where / when something is done / happens?
- What is the object of the description?
- Where and when the picture is taken?
- What is the source of the purchased material or archive copies (and restrictions on use)?

Categorisation of Programming for Content Description

The level of the description depends on the re-use value, as well as on the genre and type of the programme. In the guidelines, the programmes are divided into 5 groups:

Group A includes news and sport programmes, as well as current affairs and nature programmes with re-use rights. The content of the programmes is described and segmented according to the programme type. The extent of the content information depends on the footage, the subject and the rights of use of the programme.

Group B: Part of the images can be re-used, part may be subject to a charge. Programmes include parts of news, current affairs, music, factual and sport programmes. The content of the programmes is described more approximately than the content of the A-group programmes, because their re-use is restricted (e.g. rights to use). The content is segmented according to the programme type.

Group C: Minimal insert-use, re-use is restricted. Programmes: parts of factual, discussion and entertainment programmes. In this case, the content of the programmes is described briefly, because the re-use is limited (e.g. restrictions on use, protection of privacy etc.).

Group D includes drama and children's programmes as well as domestic purchased programmes. Programmes in E-group are films, series and other purchased programmes of foreign origin. The content of the programmes in groups D and E are not described, since their use is limited to one or a few presentations.

The guidelines assist Yle staff with the CD creation process by giving examples of programmes in each of the groups. In particular, the following programmes used in our data analysis are cited:

- Group A: *News*, *MOT* ('QED'), *Närbild* ('Close-up')
- Group B: *Puoli_seitsemän* 'Half past seven'
- Group C: *Strömsö*

According to one archive editor, the other programmes in our corpus can be classified as follows:

Programme name (translation)	Category*	Notes
Silminnäkijä ('Eyewitness'):	C- or D	Dependent on the rights to use
Sohvasurffaajat ('Sofa surfers'):	C	
Ulkolinja ('International line')	B or D	Dependent on the rights to use
Sissipuutarhurit ('Guerilla gardeners')	C	Right to use but visual content gives no reason for exact description; private identifiable persons in the images restrict the right of use vs. milieu, public figures, situations etc.
Tekijänä ('Made by')	C	
To Nightwish with Love	D	No description of the visual image; mentioning the subject of the clip is enough

*Many of the programmes above can sometimes be classified to group B, *Ulkolinja* even to A.

Figure 36: Classification of Programmes

During the interview, one archive editor clarified the question of classification:

The level of documentation depends on the rights of use, the genre and the quality of the images. One programme title can be differently documented depending on the image content. The categorization is more or less theoretical, and the categories are not, for example, groups or statuses written down in some systems. In practice, the content description of one programme title is usually done in the same way in the production, even if there are variations in rights and contents. On the other hand, the descriptions may vary according to the describer. The instructions and guidance of the archive aim at concretizing what kind of content description best serves the re-use. Documentation categories serve as the guiding principle behind this work.

It can be concluded from this valuable insight that categorisation and the corresponding protocols for content description are neither rigid nor uniformly applied, but rather that each operative judges the material in the programme on its own merits and applies the guidelines using common sense and intuition. While this may be the best way to operate on a practical

basis, it makes the introduction of automated systems which are, by their nature rule-bound, rather more problematic.

One key area of concern, however, is the widespread dangers inherent in using third-party materials which have been not only acquired elsewhere and edited into Yle's own original material, but for which single-right of use licensing (possibly with additional restrictions such as territorial limits) apply. These have to be marked in the CDs so that they are not treated as copyright-cleared, especially where they occur in a largely Yle owned footage. One editor summarises this and alternative scenarios:

If the footage of the programme is (mainly) such that its use in other programmes is limited due to its nature (e.g., identifiable private persons, processed image), it is generally briefly described, according to the guidelines, for example only the names (and the action) are mentioned, processed images may not be mentioned at all. This kind of footage with limited re-use but with no specific contractual restrictions on use is usually categorized into the C-group. Contractually restricted images (e.g. images from external sources) must be marked so that they are not (accidentally) used in other programmes - their content is only described to the extent that the sequence can be recognized; the most important thing to describe is the source of the image and the restriction on use.

The analysis of our YLE Programme Corpus ('YPC') revealed that the level of description did not always correspond to the given categories, for example *To Nightwish with Love* (category D), which was described in a very granular fashion. From the production coordinators' point of view, categorisation perceptions are different. According to them, the level of description depends on the genre and the type of the programme. Further, if the describer is part of the production team, they have a thorough knowledge of the material captured and therefore know what is most relevant to the content description. On the other hand, if they do not have any information about the shooting location, this information cannot be applied, and the description will be more general. Generally speaking, news and documentaries are described in greater detail than entertainment programmes, and magazine shows (programmes with several topics) contain more micro-descriptions than programmes with only one subject ('single topic programme')⁵. Also programmes belonging to genres which are normally not described (e.g. children's programmes) will still be described where there are images which are thought to be useful later on. Regarding licensing, a production coordinator commented that it is always possible to pay for reuse, if the footage is important and would otherwise be restricted. In addition, the re-use rights may later change. Accordingly, a more extensive description of the programmes could be useful than what is now possible regarding the resources.

Search and Retrieval: For search purposes an ideal content description does not contain too much information: a rich description would produce 'noise' that is irrelevant to the search

⁵ Yle uses both 'subject' and 'topic'; the difference is not clear.

results. CD should be general, not too concrete. It should consist of sentences, not single words (which are closer to metadata), for example *a tractor* is not enough; the description should include the action (what the tractor is doing in the image) and the location (*a tractor moves in the field*) (or: *a tractor in the field in sunny autumn weather*).

According to the guidelines, the description may contain additional epithets with specific information in brackets aiming to facilitate the search and selection of images in the archive. Typical examples are information of origin, restrictions of use, shooting time and place, technical information of image and voice, for example *inside/ outside (the Parliament Building (inside/outside))*; *anonymous image, close, speeded-up image, text on [the image or interview], mute*.

In the description of cities, nature and buildings (from the outside) it is recommended to mention the season or the time of day. If it is different from the transmission date or the programme is shot in different seasons, this is especially important. Examples: *(spring)*, *(night)*. After description of a landscape general descriptions can be added: *(rural landscape)*, *(sea view)*, *(city scenery)*, *(lake landscape in Finland)*.

An interesting question is the selection of actions which should be described. One production coordinator noted that it is unnecessary to describe very common actions, for example a reporter browsing through his papers or webpage – who wants to use such footage? Another operative described how she selects things that could interest somebody.

These subjective factors impacting the selection of elements to be content-described once again raises a number of issues relevant to automation. Firstly, if selection is left to the operative without reference to binding guidelines, there is no guarantee that one person's idea of saliency matches the next person's retrieval needs. Referring to the example above, images of a report browsing through a webpage might be highly relevant to accompany a news report on journalists hacking websites for private data. Secondly, it is nigh on impossible to describe a workflow (on paper) that would define the CD process in such a way as to build an algorithm which automates part or all of the process, when saliency is personal and unbounded.

In this regard, a common theme is starting to emerge from the two strands of investigation within WP5 (automated captions for views and automation for content retrieval): saliency in the context of narrative, multimodal storytelling is a highly subjective, intuitive and nebulous affair, often becoming a matter of disagreement between human beings. Training a computer to become 'saliency-savvy' when there are so many variables in play, is at the very heart of the AI computer vision challenge – irrespective of the precise nature of the task in hand.

4 Analysis of Content Descriptions

This section presents an analysis of two extracts from *MOT* 'QED' (group A).

In the following example, the series is first briefly described, and sample segments from the programme are analysed in detail. In order to illustrate the faithfulness of the CD, the description of the visual content is presented in tables: the right column shows the content description made by Yle, the left column our own detailed descriptions of the shots. Finally, main elements and features of the CD of the programme are outlined focusing on patterns, linguistic features, coherence and narrative elements.

MOT ('QED')

The programme *MOT* is classified as "current affairs" and its subclass as "report". Each *MOT* programme deals with one topic only.

The production coordinator, a member of the production team, explained that all interviews which take place on the show are transcribed, and the transcriptions are produced by a group of six persons who have done the job for a long time; for example, one of them since the beginning of the programme in 1996: *"They have learned the process so that they don't transcribe useless speech ('this kind of blabbering can't end up on screen')"*. The producer first selects the utterances which are used in the script and the scripts are subsequently published on the Internet. The production coordinator uses the script in the description of the content (e.g. names and titles, language code, places). The script is, however, not relevant for the description of the visual images: here it is important to think which key words will be used when searching images: *"If somebody in the picture is sitting in a car, it does not matter whether he is coming or going. This fact goes with the speech in the programme, but when you need an image it's purpose of use will change anyway."* Again, the coordinator appears to be referencing the nebulous qualities of narrative saliency.

The title of this example *MOT* programme MEDIA_2009_00019261 translates literally as: 'In the journey with a people smuggler' (English title: At a smuggler's mercy) and has a duration of 00:29:19:02 (hh:mm:ss:ff). The description is divided into 7 segments, the first segment consisting only of the *MOT* logo.

Extract 1: Segment 2 (3)⁶, 00:00:12,124 --> 00:06:48,560

	Shot description	Content description
1	A people smuggler seen from behind is sitting at a window, smoking and looking out. Outside only a tin roof and trees to be seen. He talks.	*people smuggler looks out of a window. Interv. people smuggler 1 **(XX+) (anonymous image).

⁶ In brackets Yle's segment number when it does not match with the chronological order.

2	Text on black screen: <i>Ihmissalakuljettajan matkassa</i> 'In the journey with a people smuggler'.	Title: 'In the journey with a people smuggler'
3	It's dark. People are sitting on the ground.	People are sitting quietly on the ground (night).
4	Men walking with packs on their back.	A group of people are walking in the dark (night).
5	Two men lying.	
6	A boat comes ashore. Two men are waiting on the shore.	A boat comes ashore, people on the shore (night).
7	Flying Turkish flag. Text: Istanbul in July 2009. Minaret.	The Turkish flag (text on it).
8	Images of the city.	Istanbul, the city (long shot).
9	Traffic on the street.	Traffic in Istanbul city centre.
10	The Bosphorus.	
11	The smuggler.	Interv. people smuggler 1 (XX+) (anonymous image).
12	Traffic in the city. One house is badly run-down.	Traffic in Istanbul.
13	The smuggler.	Interv. people smuggler (XX+) (anonymous image).
14	Traffic on the streets. The Turkish flag hangs from a window. Minaret.	City scenes from Istanbul, people at the market.
15	People are passing between market stalls. A man stops to watch sports shoes. He tries on shoes.	Men walk between market stalls. (A) man is trying on shoes in a stall.
16	The smuggler.	Interv. people smuggler (XX+) (anonymous image).
17	The man who is buying shoes has a plastic bag in his hand. The seller takes	(The) man pays (the) seller for the shoes.

	banknotes out of the pocket and gives some to the man, who then leaves the shop.	
18	The smuggler.	Interv. people smuggler (KU+) (anonymous)

Extract 1: MOT [MEDIA_2009_00019261]

* Finnish does not have articles, therefore articles are left out in the translation above: The variation between *a* and *the* would make the identity/continuation of the referents explicit, whereas the original may leave it open. In Finnish, definiteness is indicated by other means, such as case or demonstrative pronouns.

** (XX) is a language code meaning 'speaks an unknown language'. (KU) means 'Kurdish'; + = the person is visible (– = not visible, telephone interview).

The table illustrates that the content description is relatively faithful to the imagery. However, not every shot or detail is described, nor does it need to be. The length and quality of the image are general reasons for not describing, but still the question of selection remains. For example, from the scene 'buying shoes' (lines 15 and 17) the actions 'trying on shoes' and 'paying for the shoes' are described but not 'the man holds a plastic bag in his hand' which might be considered as irrelevant or less useful for the search or re-use purpose.

The description "pays for (*sic*) the shoes" contains an explication: In the picture, we can only see two men handling money and one of them (the seller) giving notes to the other. How do we know that the one is a seller and the other is paying him for the shoes? The describer uses his/her schematic knowledge (schema 'buying') and links the events in the shots together into a story. In English the choice of the article would make clear that *man* in the buying schema refers to the same person. In Finnish CD the identity of the reference is open. A production coordinator commented on this: "The fact that 'man' refers to the same person in the sentences (*man is trying on shoes ...*, *man pays the shoes*) is irrelevant to the CD.

The extract shows variation in the description of similar images. For example, the same kind of urban scenery is described with different words: *Traffic in Istanbul*; *City scenes from Istanbul*; *Buildings and traffic in Istanbul*. According to the guidelines a pictorial motif in a segment is described only once, even if the footage is cut into several parts. This applies to interviews as well, however, in *Extract 1* the interview with the smuggler is mentioned every time (5 in all).

On the basis of the whole corpus, mood is not a usual element in Yle's content description. *Extract 1* shows a rare example of description of mood: *People are sitting **quietly** on the ground.* (3)

Descriptions in brackets are additional information which is given according to the guidelines describing the time of day (*night*) and containing special information of the image (*anonymous image*).

Extract 2. Segment 4 (5): 00:12:47,680 --> 00:16:26,761

	Shot description	Content description
1	A map of northern Turkey showing a route from Istanbul to Enez.	Map: Route from Istanbul towards the Greek border.
2	People carrying backpacks and plastic bags walk in the woods. They stop for a rest. Some men are watching the environment. They continue on their way, hurry and stop to wait in between.	Persons being smuggled walk in the woods with their goods. (The) group on a rest break. (The) group is sitting on the ground and watching the environment. (The) group goes in the middle of the woods.
3	It is night. People are sitting on the ground.	Persons being smuggled are sitting on the ground and waiting (night).
4	A boy cries and a woman comfort him.	(A) boy cries, (a) woman tries to comfort him (night).
5	Two men are lying.	
6	A woman prays.	(A) woman prays,
7	The group continues their journey.	(the) group continues the way in the woods (night).
8	The map shows the border between Turkey and Greece at the Mediterranean coast. The route goes across the sea from Enez to Greece.	Map: Proceeding of (the) group on the map at the border between Turkey and Greece.
9	Men are sitting on the shore. A boat arrives. Some men get out of it onshore.	The men are sitting on the ground. (A) boat arrives onshore. The men come on the shore (night).

Extract 2: MOT [MEDIA_2009_00019261]

This segment can be understood as a narrative about the journey of the people who are smuggled. Narrative cues can be found in a number of markers: continuation of referents, description of successive actions (e.g. *wait – continue*), definite nouns (*the men*) (in Finnish the subject *miehet* in the nominative case, the indefinite form would be partitive *miehiä*).

The description is not only based on the visual image on screen but has also required interpretation and/or information about the whole film:

*Persons being smuggled are sitting on the ground and **waiting*** (3)

*A boy cries, a woman **tries** to comfort him.* (4)

The segment shows a rare example of a pronoun: *A boy cries, a woman tries to comfort **him***. It is to be noted that the pronoun occurs in the same sentence as its antecedent.

As the extracts show, the description renders the visual information at a relatively fine, granular level. One explanation is that *MOT* is well resourced, and the production coordinator can spend more time on the description than may be available for other programmes.

In the analysis of the CD of the whole programme three main types of descriptions were found (cf. the key elements in Deliverable 5.1 *Multimodal Annotation of Described video*):

1. Action scenes where following elements occur: character, action, object, location; the type 'character + action + location' being most frequent.
2. Landscapes: short descriptions of the view (*traffic*) and/or name of the place (*City of Istanbul (urban landscape)*; *Buildings and traffic (street scene) in Istanbul*)
3. Interviews: *Haast* (abbreviation of 'interviewee, being interviewed') + person + language code

Typically, the linguistic features of the CD include relatively simple syntax. Different elements in an image and successive or simultaneous actions can be described in one sentence and separated with a comma (or conjunction *and*): *A woman prays, the group continue their way in the woods (night)* (Ex. 2: 6-7).

Although visual information is described, speech offers cues for the description:

The men are sitting on a couch in the smugglers' flat. – 'smugglers' flat' is mentioned in the reporter's speech.

The men are sitting on the floor of a van on the way towards the Greek border. – Reporter: 'The journey to the Greek border has begun'.

The CD contains several examples of explication, for example it links two separate images together in the description '*The men are watching TV*'. In one shot we can see men sitting and looking at something, in another shot we see a television but not the watchers.

Lexical variation occurs, for example the same figures in the programme are characterized as *pakolaiset* 'refugees', *joukko* 'group', *salakuljetettavat* 'persons being smuggled'. The production coordinator was challenged about this variation and the implications for cohesion. Where there are repetitions, the coordinator uses copy and paste to reduce typing efforts; however, if this continues, the connection between the elements becomes irrelevant and is

not described. Since *MOT* has only one topic, retrieval of one passage almost invariably leads to retrieval of other segments in the same vein, such that there is minimal impact on issues of narrative coherence where multiple terms are used for the same concept. However, the description of the people as refugees or persons being smuggled requires knowledge of the whole programme, for example, the last image is described with the following sentence: *A/the group of refugees walks in the dark (night picture)*. On screen, figures are walking but how can one know that they are refugees? (The prior image shows Ali in the smuggler's flat.)

CD is assumed to be less narrative than audio description. However, since *MOT* has only one topic and the example programme tells the story of refugees being smuggled from Istanbul to Athens, the whole CD can be read as a story. One character, Ali Hakmat Baker, is interviewed in segment 3 (*Interv Ali Hakmat Baker (KU+)*) and is the main figure in segments 5 and 6. Segment 5 contains two references to Ali: At the beginning in the description of an interview (as in seg. 3), and in the last sentence (*Ali walks with the group.*). Between these descriptions, actions of the group are described. Thus, *group* in the sentence *Ali walks with the group* can be inferred to refer to the same group (and *Ali* to Ali Hakmat).

5 Findings on the whole corpus

The above examples provide an outline of the methodology applied in the creation of the content descriptions and the types of considerations that guided those engaged in CD creation. This following section presents a more reflective analysis of all eighteen programmes studied for the purposes of our investigation. The first section focuses on the structure of CD, level of granularity, description of speech and cohesion. The second part concentrates on the main linguistic features (grammar and lexis) found in the corpus. Interviewees' comments are included, as far as these issues were dealt with in the discussions.

5.1 Structure

In the earlier results produced from WP5, the team compiled a list of narrative building blocks found in all film material (described in Deliverable 5.1) which they called 'key elements'. These comprised: characters, actions, salient objects, locations and mood. The same elements can be found in CDs with the exception of mood which in the corpus is described only in few cases (e.g. *The Finnish children **seem to be tired** of the job; People are sitting **quietly** on the ground.*). One production coordinator explained that she describes emotions if they are conveyed by the image, for example: a crying person, laughing children; happy kids are jumping in puddles in driving rain. An archive editor emphasised the objectivity: for example, 'touched' is acceptable, but 'furious' would not be permitted because it is making a strong judgement call about the expressed emotion.

In addition to the key elements, there are other elements in CD called "additional information", contained in brackets. This is information regarding filmic characteristics and

camera movements (e.g. close-up, long shot; processed image, anonymous image) or image composition (outside, inside; winter). In practice, the describer can put any type of additional information in brackets, for example: *noticeboard (ugly)*. In the search, however, it does not make any difference whether the element is bracketed or not; it only indicates that the element in brackets is additional information. An archive editor commented on the description (*ugly*): “It may indicate that this is an interpretation, but it is not according to the guidelines.” The description can also include diverse elements which refer to time: *evening scene; dawn; morning traffic; ilta pimenee* (‘evening gets dark’) ‘it’s getting dark’.

In CD, the action is often accompanied with a **direction** or a **goal**. The description of these elements is not based on the image but on the describer’s knowledge of the content.:

Direction: *In the minibus on the way back; The women leave for lunch.*

Goal: *A group of men is carrying colourful brushes for sale.*

In the picture, men walking in woods and carrying colourful things can be seen, but not now or later that they are selling these things.



An archive editor commented on the description ‘A goes to ask B’s advice’ and that ‘asks advice’ is unnecessary since ‘A goes to B’ or ‘A and B’ would be enough. She emphasized that it is unnecessary to tell a story; finding the programme is most important. Similarly, a programme coordinator said that she wouldn’t put ‘where he is going’, because you can’t see it in the image; neither ‘why he is walking’, but ‘how’, for example, ‘scuffing’, because somebody could seek ‘a person hurrying forward’.

5.2 Level of granularity

An interesting question is how concrete and detailed the description should be. The description can catch concrete visual cues and happen at a very basic level as in the following example:

Akuna Bay National Park in Sydney (Australia). River and stones in rain. Ashley Redman (EN+) stands/is standing in the middle of the river and speaks to the camera. Ashley sits/is sitting on the bed. Ashley lies/is lying on the floor. Ashley sits on the floor next to the window and speaks. (To Nightwish with Love)

An archive editor commented on the description *Ashley stands -- Ashley speaks -- Ashley sits - Ashley lies* –that it is too detailed; ‘Ashley sits in her room’ would be enough. According to her, it is very unlikely that someone would search for an image of Ashley lying on the floor. In addition, the re-use of images showing persons is often restricted for various reasons

(including personal privacy, respect for non-public figures, and ensuring video clips are not re-used in a way that might re-purpose original comments out of context etc.).

Furthermore, the description must not be too detailed. One archive editor stated that details are not described, unless they are especially relevant, the shot is long (10 seconds) and the image of the object is very good. For example, if there are several images of flowers, there is no need to describe every flower. If there is illustration of ship industry, “ship welding” might be a too detailed. The description should be general, for example ‘soldiers in Syria’, not: ‘soldiers come from tank, take out weapons and begin to shoot’. This principle is one reason for the summarizing descriptions.

One-word descriptions are for search purposes problematic, because the context and the relation to the action or event are missing; for example *dog* is too wide; whereas the preference would be to use: *a small white dog* (+ sentence). The description of species of animals, birds and plants depends on the describer’s knowledge. One production coordinator said, that she always puts the dog breed, if she knows it. (This can be problematic too if the term is unknown to the archive user, e.g. *Dandie Dinmont Terrier*). She also explained that if the image shows mountain scenery, she does not write ‘mountain scenery’ but for example ‘the Swiss Alps’. ‘Atmospheric episode’ or ‘Idyll’ does not say anything, the description should explicate what is in the image.

According to an archive editor, the context, the **action** on screen should be verbalized; the choice of a term may be difficult, therefore a sentence is better. For example, ‘Breads on the production line’ is not good, because the verb is missing, a better description would be ‘Breads on a conveyor belt, baked rye breads are bagged.’

According to the guidelines, short images (shorter than 10 seconds) forming a coherent whole can be described by a general characterization of the visual contents, for example *Street scenes of Addis Abeba, Ethiopia*. In the analysed data summarizing descriptions is frequent. Summaries contain information of a longer sequence or the whole film which need not be visible in the image. The following description of a two-minute segment illustrates this:

*Veikkola school in Kirkkonummi (outside, winter). Teacher Satu Kivinen and the fifth-graders of Veikkola school **start the solar cooker project** in the classroom. (Dokumenttiprojekti)*

First, there is a picture of a building. Next, the film shows children and a woman in a classroom. In the sequence described above we see following: The woman hands out notebooks (English subtitles: *I'm handing out notebooks which you'll always use as we do these tasks. Write your own name and "Solar cooker" on it.*). She is talking, the children raise their hands, speak and look forward. The woman points at a map on the wall. The children are drawing on the notebooks, the woman goes between them, leans over them, and speaks. The description *start a project* refers to the beginning of a work that is not concrete, visible; instead, the

characters are talking and drawing. Nor are any solar cookers visible. The description explicates a beginning and entails the implicature of a story with beginning and ending. The description of the next segment is also a kind of summary and as such general and abstract:

The pupils build a solar cooker out of a cardboard box in the school workshop. Eemi, Roope and Henna among others building the cooker.

On screen we can see children acting with different objects: a big box, glue, small metal pieces and screwdrivers, but from the visual content only it is not possible to draw, what they are doing. The three children mentioned by name are key figures in the story.

A production coordinator said that the longer the unit, the more compact the description; it is easier to tell more about one picture than about a sequence of events. “For example, I could write ‘N is sitting in a park’, but then the question arises: ‘Does he sit on a bench, on the ground...’” On the other hand, summaries may contain more complex sentences than detailed descriptions.

5.3 Description of speech

According to the guidelines, in most cases it is not necessary to specify the **topic** of an interview (if it can be inferred from the context or the subject of the programme or insert). Further, it is recommended to describe it on the subject strata. Accordingly, a short description of the interviewee suffices, for example: ‘Nurse Regina Lange (SA+)’. (The language code SA means that the person speaks German, + refers to her visibility in the image.) If the content of the interview is relevant to the search, it is briefly specified with a few words. An archive editor said that if there is something exceptional in the **dialogue**, it can be described. The problem is that often we do not know until much later whether an utterance will be long lived and frequently quoted.

The description varies according to the programme. In “Parliament question time”, ‘X speaks’ does not suffice, but for example: ‘X speaks of the social welfare and healthcare reform’.

In the analysed data, the topic of the discussion is described in various ways:

- *N1 N2 / N1 discusses with N2*
- *N talks about X: N talks about the status and abuse of women*
- *N tells that there is an abused woman in the changing room. (Sohvasurffaajat)*
- *Interv. NN (EN+), the baby hatches prevent leaving new-borns in rubbish bins. (Silminnäkiä)*
- *The Finnish children comment on their experiences on the first working day. (Dokumenttiprojekti)*
- *In the documentation of a discussion, the host’s questions are quoted: The discussion continues. How is hostility against strangers effectively resisted? (Eurovaalit)*

- *Reporter Valkeeniemi listens the message of the alarm phone and repeats it in German.* (Silminnäkijä)
- *N presents a hatch and instructions.* (Silminnäkijä)
-

Close transcriptions of the audio with some omissions also occur, although these are against the guidelines. In the following example words which are identical in the CD and Finnish subtitle are marked.

- CD: *Haast mies (ES+): **läksimme maasta, sillä Venezuelassa vallitsee huono taloudellinen ja poliittinen tilanne.***
'Interv (ES+): we left the country, because in Venezuela the economic and political situation is bad.'
- Finnish subtitle: **Lähdimme maasta** parempien työntekomahdollisuuksien perässä. **Venezuelassa vallitsee huono taloudellinen ja poliittinen tilanne.** Oli pakko lähteä. Olen ammatiltani keittiömestari, — eikä minulla ole, mistä valmistaa ruokaa. Aineksia ei ole.
- English subtitle: **We left the country** to seek better employment opportunities. **The economic and political situation is poor in Venezuela.** It was necessary to leave. I'm a chef — but I have no food to prepare. There is no food. (Ulkolinja)

5.4 Cohesion

CD is not read as a text, since its main function is to serve as a resource to complete an archive search. The description *refers* to an image or sequence and should lead to it; it does not *replace* the visual content (cf. AD). The recipient seeks images or film footage for re-use in new programmes and employs short search terms which should correspond to terms used in the CD. Since single sentences or phrases are needed, it could be assumed that the creation of coherence or cohesion would not play any role in the CD. Also the interviewees emphasized the preference of short unconnected sentences, which the analysis showed to be most frequent (cf. more in the section 5.5). Another explanation for the lack of cohesion is that copy-paste-method is often used in description of reoccurring images.

Chronology creates coherence, too. In CD, programmes and segments are described chronologically, but each visual topic is described only once in a segment, even if it is edited into multiple sequences within the segment. Accordingly, the descriptions within a segment do not necessarily have to be in chronological order.

Despite the irrelevance of cohesion in CD, cohesive descriptions of segments constituting a narrative could be found. In the following example, cohesive ties include use of synonyms in references to same characters (*the Finnish group, the Finns, the children*) or entities (*school and the nature club, the class*) as well as elements of schematic description of a travel (*arrives – on the way back*).

Evening scenery from Addis Abeba (speeded up image). Yemane Birha School in Addis Ababa, Ethiopia. The Finnish group arrives at the school and the nature club. The Finns present themselves in front of the class. On the way back, the children comment on their first impressions. (Dokumenttiprojekti)

5.5 Linguistic features

Grammar

Syntax: According to the guidelines, the description should be written in complete sentences. In the analysed cases, the syntax is usually simple, typical structure being: subject + predicate + (object) + (location). The verb (copula) can be left out (as well as an identical subject within a segment). A production coordinator said: “*One sentence one thing. Main clauses, no subordinate clauses.*” Another suggested that she tries to describe as clearly as possible, for example ‘Tapio Rautavaara [= a late singer] is sitting on a park bench with a backpack on his knees’. When asked “Is the backpack important?” she answered: “It’s part of the atmosphere.”

Examples of simple syntax:

- The boat arrives to the beach, people on the beach. (MOT)
- Men are watching television. (MOT)
- Parikka [= person’s name] at home. Packs her suitcase. (Tekijänä)
- Texts (voice: quotations are read). (Silminnäkijä)

Descriptions of single words without verbs (referring to an action) also occur:

- Baby hatch (inside), security camera, documents. (Silminnäkijä)
- A tractor. -- A dog. A horse. Flowing river. (Närbild)
- Pedestrian street of Kouvola, people on the street, two old women at a market stall, people on the benches. Two men are talking, empty business space, “for rent” note in a window. (Yle Uutiset Kaakkois-Suomi)

The action can be uttered by nominalization or with a noun: Instead of *Parikka is discussing scarfs* the situation is described with *Parikka in a scarf meeting*. When the active verb is nominalized in Finnish, the object of the action is expressed with attributive genitive: *kylttien kirjoitus* ‘signs+GEN writing’. Further examples: *tarvikkeiden sosialisointia* ‘socialization of supplies’; also: ‘people in their everyday chore’ (instead of cooking etc.). These kinds of descriptions are used to summarize the content. Nominalizations and the use of non-finite verb forms can lead to very complex sentences in which the description is condensed as in the following example: *autonsa pysäyttänyt mies* ‘car+ACC+POSS stopped man’ ‘a man who has stopped his car’.

Tense: Descriptions are usually written in the present tense as they describe actions or scenes which are in the very moment visible on screen. The following example shows an exception:

*A and B show their own press cuttings about dreams and visions which **they are going to fix** on the inspiration board **made** by them[selves] in Villa Strömsö's crafting room. (Strömsö).* Finnish grammar does not have any special form for the future tense. The use of the verb *tulla* 'come' can be used as a kind of auxiliary (as in the Finnish description), but it is not regarded as good style. The form 'made by them[selves]' refers to an action prior to forthcoming action (fixing) but after the present action (showing).

Voice: Sentences are mostly in active voice, but also passive occurs, for example: *The work is completed; A car is driven; Children are pushed in strollers.*

Aspect: In some cases, aspect of the action is expressed by verbs e.g. *lähtevät matkaan* 'leave (for a journey)', *jatkaa matkaa* 'continue (the journey)', *yrittää lohduttaa* 'try to comfort'. These examples come from the CD of *MOT* and indicate its narrative character. The progressive aspect is mostly not marked in Finnish but expressed with the present simple (e.g. *ruokailee* 'eats') but occurs as (*ruokailemassa* eating'). In Finnish, aspect can also be expressed with cases (corresponding to prepositions in English), e.g. *kävelee kadulla/katua* 'walks on the street / along the street', *rannalla/rantaa pitkin* 'on the beach / along the beach'.

Complexity of constituents: Nouns are often specified with genitive case, e.g. *salakuljettajien asunnossa* 'in the **smugglers'** flat', *Ateenan ravintolassa* 'in a restaurant of Athena'; also: (*istuu*) *sängyn reunalla* (vs. *sängyllä*) (sits) on the edge of the bed' (vs. on the bed). The key element 'localisation' may be quite complex as in the description (*piirtävät*) *asfaltille Hinthaaan kaupan edessä* '(draw) on asphalt in the front of a shop of Hinthaa'. Characterization with adjective attributes is less frequent but occurs: *A plastic bag flies in the rocky landscape of Petra in Jordan.*

Lexis

In the lexical analysis, the corpus tool *Sketch Engine* was used. The aim was to find out which verbs are used in description of actions and which nouns in description of characters and detect variation. The whole corpus (18 programmes) consists of 7,805 tokens. Many CDs contain transcriptions or quotations of speech which are not descriptions of the visual information, therefore words in these sequences were excluded as well as the language codes. Only topics of the talk were included in such sentences as 'N tells about X'). The cleaned data consists of 5,482 tokens and 4,145 word tokens. Compared to the English corpora presented in Section 3.6, the type-token ratio 0.526 is very high due to the rich Finnish morphology. If instead of types (unique words) lemmas (a word with its all forms) are calculated, the ratio is still high: 0.442. This does not only indicate lexical richness, but can also be explained by identification of persons and places with names. In the following, the numbers in brackets refer to real numbers of the lemmas.

Actions

The total number of verbs amounts to 513 including 200 different verbs. As many as 122 verbs occur only once which indicates great variation. Most frequent verbs are *kävellä* 'walk' (27), *istua* 'sit' (25), *kertoa* 'tell' (25). In addition to 'walk' motion is frequently described with *saapua* 'arrive' (9) and its antonym *lähteä* 'leave', 'go' (7), whereas the common verbs *tulla* 'come' (4) and *mennä* 'go' (3) are used less frequently in the corpus. A programme coordinator commented on the choice of a verb that she would rather describe 'walks slowly/fast' than 'runs', or if the person walks with a limp, then 'limps', but without much variation. This is likely to be a means of streamlining lexicon, since word-searches do not operate on synonymity, and a search for 'walking' would not return 'running', whereas 'walks slowly or walks fast' would still reap results.

Images presenting static characters can be described with 'character + location', thus avoiding use of a verb. Besides the verb *istua* 'sit', the verbs *seistä* 'stand' and *maata* 'lie' are used, but much less often (6 occurrences of each). The frequency of the verb *kertoa* 'tell' indicates the relevance of speech; after the conjunction *ja* 'and' the most frequent word in the corpus is the abbreviation *haast* 'interv' (111) which refers to 'interviewee/being interviewed'. Other verbs describing talk are *puhua* 'speak', 'talk' (14), *lukea* 'read' (12), *keskustella* 'discuss' (8), *esitellä* 'present' (8).

Frequent verbs describing other kinds of actions *näyttää* 'show' (14), *soittaa* 'play' (12) and the general verb *tehdä* 'make' (10) which varies with many other verbs e.g. *valmistaa* 'prepare' (11) and *rakentaa* 'build' (6). The corpus contains various kinds of actions, and accordingly the descriptions varies, e.g. *avata* 'open' (5), *maistella* 'taste' (5), *katsella* 'watch' (5), *katsoa* 'look' (4). The use of different word forms increases the variation. For example, in Finnish it is possible to build a frequentative indicating repeated action: *istua* 'sit' – *istuskella* 'sit around', *seistä/seisoa* – *seisokella* 'hang around', both occur in the corpus. Many verbs implicate interpretation or knowledge of the context (e.g. *luonnostella* 'sketch', *pelleillä* 'fool around', *testata* 'test' or *viimeistellä* 'finish').

Characters

Characters are typically identified and described by name and title as instructed in the guidelines. These descriptions are here excluded and only descriptions with common nouns are analysed. The total number of nouns describing persons amounts to 250 including 78 different words, most frequent being *mies* 'man' (35), *ihminen* 'human/people' (30), *lapsi* 'child' (16), *joukko* 'group' (15) (also the synonym *ryhmä* occurs), *nainen* 'woman' (11). Often more specific nouns are used, for example: *suomalaiset* 'Finns' (7) or occasionally such as *kerjäläinen* 'beggar', *naislääkäri* 'woman doctor', *turisti* 'tourist'. Different descriptions of the same characters produce lexical variation and increase cohesion, for example: *pakolaiset* 'refugees', *joukko* 'group', *salakuljetettavat* 'smuggled persons' or: *lapset* 'children', *koululaiset* 'pupils', *oppilaat* 'students'.

As in the description of actions, the use of different common nouns shows that the description is not only based on visual cues, but also on other information (audio, text) and general knowledge (cf. the nouns ‘neighbour’ and ‘bedouin’).

6 Summary

There is great variation in the CD depending on genre, type (discussion, magazine show, one subject documentary etc.), production team, describer, language (Swedish describers may rely on Finnish subtitles, since the descriptions are made in Finnish). The variation concerns level of granularity, complexity of the structure and lexis. Depending on details, there is a continuum between close descriptions and summaries: at one end the description is very faithful describing each shot and its different frames, and at the other, it summarizes several shots. Sentences in summarizing descriptions tend to be longer and more complex. Descriptions of objects may catch details by adding attributes and adjuncts qualifying objects, characters or actions, which again lengthens and complicates sentences, or more specific terms can be used. The question is which strategy is more useful for search purposes and therefore which would be most relevant to prioritise in the machine learning context. Finally, descriptions can be classified into three broad categories: (1) actions, (2) landscape and (3) speech.

The CD should consist of simple sentences, but there is a large variation from one-word utterances to complex sentences. The describer should only describe what is visible on screen, but *many descriptions are based on the describer’s encyclopaedic and everyday knowledge and background information*, which once again has resonance with the way human beings compile content descriptions for feature film extracts (see Part A). For example, many descriptions of actions contain ‘goal’ or ‘intention’ information, although this is regarded as irrelevant to the search. The question arises how far it is possible to describe images without drawing inferences. The lexical breath is one indication of building on different kinds of knowledge when describing visual images: human-made CD contains much variation, although it is not required or even desirable.

Content description (at Yle and at the present stage) differs in many ways from audio description which is designed to enhance accessibility for sight-impaired audiences, whereas the purpose of CD is to enable the re-use of the images and efficient search in the archive. Ideally, CD contains isolated sentences which refer to key images. Although narrative elements do sporadically occur, their occurrence is unintentional and irrelevant. Still, some descriptions of segments as well as separate sentences can be read as a story although perhaps this might be classified as purely coincidental. In contrast to AD, in CD cohesion is irrelevant as well, and continuation is usually not marked. CD is created segment by segment, and the length of a segment varies. The length is not dependent on audio track as in AD where

the description is inserted in hiatuses. In CD identification of characters and places is important, one area of overlap with audio description despite the difference in orientation. Repeated actions/images are described in a segment only once, and chronological order is not important, since in the search the user has access to the images as well. Some key elements seem to occur in AD and CD with the exception of mood which is only rarely described in CD, and not particularly common in AD. Finally, the role of speech distinguishes CD and AD: in CD talk is at most briefly described, whereas AD is part of a whole audiovisual programme with original speech content. The dissimilarities clearly demonstrate the main difference between the two types of description: CD is ancillary text, AD surrogate text, as stated in Deliverable 5.1.

Content description absorbs considerable resources which limits the extent to which it can be entertained by a commercial broadcaster, and the level of consistency and relevance to the end-user achieved where it is employed in production. Automatic captioning would increase the volume of descriptions available and therefore improve access to archive material, enhancing re-sale opportunities.

A comparison between machine generated descriptions and human descriptions of the present corpus will be the next step in the study.

PART C

Conclusions and Recommendations

The human analysis of machine-generated multimodal captions that has been undertaken during the most recent phase of the project has shown that the automatic generation of natural-language descriptions of video scenes still presents a non-trivial challenge for both the computer vision and the language-processing communities.

At the most fundamental level, object recognition in moving imagery requires considerable improvement before it can be relied upon to build meaningful narrative. Currently, basic errors are endemic in the captioning process: a desk is mistaken for a surfboard, a picture frame for a laptop computer, a woman for a man, black for red. Until these issues are resolved, higher order problems such as pinpointing saliency, establishing cohesive ties within and across sequenced storylines, and incorporating story grammar to frame narrative, are by and large redundant. We can use moving image derived machine-generated captions to construct the simplest story in the world, but if a telegraph pole is captioned as a microphone, and a house is labelled a clock, the result will invariably equate to nonsense.

To move the WP5 forward, we believe modelling multimodal content description will effect a deeper understanding of the human meaning-making process and incorporate recommendations for machine learning which go beyond current computer capabilities and extrapolate into future projects and areas of computer vision research. Furthermore, we can make a number of recommendations which relate more directly to the computer descriptions analysed to date, and which we believe may represent next steps in the MeMAD project.

(i) Feature extraction to improve accuracy of object recognition

At present, object recognition is largely dependent on the availability of training data which is not sufficiently comprehensive in either volume or breath, to provide the variety of non-iconic angles of common objects, or catalogue the range of permutations and variations in their appearance, for accurate AI-detection. Feature extraction techniques may offer the best short-term solution to improving the status quo, with the focus on modest areas of improvement (e.g. male/female identification, colour labelling). However, large-scale, high-quality, open access datasets compiled by lexically sophisticated operatives would arguably provide the most rapid improvement in captioning standards.

(ii) Sequencing

The current concentration on still images and GIFs in the training data, even where these are offered in short-burst sequences of five images across a simplistic narrative, fails to assist

with the development of neural networks that are sufficiently robust to infer cohesion between heavily themed multimodal material. One of the problems we have encountered, is that the computer model has no ‘iconic’ or ‘paradigmatic’ reference framework within which to contextualise identified objects. To this extent, most image captioning models appear to build from the ‘bottom up’, first seeking to identify objects, then (in the context of limited task challenges) attempting to ascribe a relationship between objects in the manner outlined in the ‘Visual Genome’ project (Krishna, 2017) and only thereafter attempting to describe the narrative underpinning each scene. This approach fails to deliver a human-like role in meaning-making, which operates on ‘top down’ basis as well as ‘bottom up’. For example, watching an episode of a television ‘whodunnit’ show begins with questions like ‘where are the protagonists situated?’, ‘who are the good and bad guys?’, ‘how are these people connected?’ and ‘what do I think is going to happen?’. Within this framework, we begin to look for people with problematic relationships, identify locations as potential crime scenes, and perceive objects as future weapons. The computer model does not understand ‘the rules’ of this game, and sees ‘a tree’ (not a forest), a man (not a murderer) and a gun (not a future murder weapon). Of course this is just one example, but the human mind works according to paradigms and rule-bound frameworks, accessing first the iconic before interpreting via the minutiae.

We suggest, therefore, that it might be beneficial to re-train the computer model to recognise narrative schematics or paradigms that are inherent to human understanding (the birthday party, the classroom, an office environment, a football match etc.), and applying these to draw on relevant lexical synsets in order to produced automated image captioning. The cohesive ties are present from the outset, rather than having to be established from a disparate array of objects and personae.

Improving the identification of characters between frames and scenes would allow for the application of pronouns which fail to deliver cohesion at present. If character A (male) and character B (female) are talking in shot one, and we know that both characters A and B appear again in shot two, ‘A man is talking to a woman’ (shot one) can become ‘He is (still) talking to her’ (shot two), rather than simply repeating the same captions for shot one and two. Both costume colour labelling and improved facial recognition techniques could assist in this regard.

Work undertaken at INA (see above), together with Aalto’s early testing of bounding boxes around recurring characters in moving image clips, suggests that further improvements in facial recognition should be possible within the lifetime of the project. There may also be opportunities within the pre- and post-editing processes for operators to identify key protagonists and assign character names to faces, which subsequently inform computer captioning choices. Likewise, vocal profiling has the potential either to assist with character identification or validate computer vision selections.

(iii) Linguistic Modelling

As a result of our extensive corpus analysis, we have concluded that AD is not directly comparable with MD, and alternative human-derived datasets are more helpful for training the model. As discussed, CD is a more reliable data source for the machine, but most importantly, the quality of future MDs is dependent upon a more syntactically flexible, lexically sophisticated, and coherent model for storytelling.

The re-introduction of ‘longtail’ words which originally featured in the crowdsourced captions (i.e. those featuring fewer than four times across the MS COCO/TGIF corpus but were removed for the sake of processing expediency) could improve lexical variety which was very poor in the MD results. However, we anticipate that this would not radically change the broader outlook.

Experimentation with paragraph captioning is ongoing, with more descriptive material and richer language being one early observation. However, the accuracy of this additional material remains to be tested, with first attempts appearing to be somewhat inaccurate/incorrect. Levels of narrative cohesion would seem to be the same as in earlier captioning outputs.

(iv) Quality and Ethical

The most important practical, and certainly ethical, point that emerges from the data presented here is that poor-quality MDs cannot replace human AD as a service for sight-impaired audiences, as they do not meet legal requirements for the provision of meaningful description. However, lower quality (albeit descriptively accurate) MDs may be acceptable for data retrieval purposes in commercial scenarios where certain film material lies outside the prime-resale category i.e. as a means of increasing marginal profits by re-purposing those video assets considered less valuable and therefore not currently warranting human annotation.

References

- Aafaq, N., Zulqarnain Gilani, S., Liu, W. and Mian, A. (2018) *Video Description: A Survey of Methods, Datasets and Evaluation Metrics*. URL: arXiv:1806.00186v1 [cs.CV]
- AENOR (2005) Norma UNE 153020: Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías. [The Standard UNE 153020 Audio description for visually impaired people. Requirements for audio description and for the production of audio guides]. Madrid: AENOR. URL: <https://www.aenor.com/normas-y-libros/buscador-de-normas/UNE?c=N0032787>.
- Braun, S. (2011) 'Creating Coherence in Audio Description'. *Meta*, 56 (3), pp. 645-662.
- Braun, S. (2016) 'The Importance of Being Relevant?'. *Target*, 28 (2), pp. 302-313.
- Braun, S. and Starr, K. (forthc.) 'Comparing human and automated approaches to visual storytelling.' In S. Braun & K. Starr (eds.) *Innovation in Audio Description Research*. London: Routledge.
- Chen, X., Fang, H., Lin, T-Y., Vedantam, R., Gupta, S., Dollár, P. and Zitnick, C. L. (2015) *Microsoft COCO Captions: Data Collection and Evaluation Server*. arXiv:1504.00325v2 [cs.CV].
- Diffalah, D., Filatova, E., and Ipeirotis, P. (2018) 'Demographics and Dynamics of Mechanical Turk Workers'. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, February 5th – 9th, Marina Del Rey, USA.
- Doukhan, D., Poels, G., Rezgui, Z. and Carrière, J. (2018) 'Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach'. *VIEW Journal of European Television History and Culture*, 7 (14), pp.103-122.
- Ekman, P. and Friesen, W. (2003) *Unmasking the Face*. Cambridge: Malor Books.
- Fresno, N., Castellà, J. and Soler Vilageliu, O. (2014) 'Less is more. Effects of the amount of information and its presentation in the recall and reception of audio described characters'. *International Journal of Sciences: Basic and Applied Research*, 14 (2), pp. 169-196.
- Han, L., Roitero, K., Gadiraju, U. (2019) 'All those wasted hours: On task abandonment in crowdsourcing'. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 11th - 15th February, Melbourne, Australia. ACM , pp. 321-329.
- Huang, T-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C.L., Parikh, D., Vanderwende, L., Galley, M. and Mitchell, M. (2016) 'Visual Storytelling'. In *Proceedings of the NAACL-HLT*, June 12th-17th, San Diego, USA.
- Hyks, V. (2005) *Audio Description and Translation: Two Related but Different Skills*. *Translating Today*, 4, pp. 6-8.
- Ibanez, A. (2010) 'Evaluation criteria and film narrative. A frame to teaching relevance in audio description'. *Perspectives: Studies in Translatology*, 18 (3), pp. 143-153.
- Independent Television Commission (2000) Guidance on Standards for Audio Description. URL: audiodescription.co.uk/uploads/general/itcguide_sds_audio_desc_word3.pdf.

- Jimenez and Seibel (2012) 'Multisemiotic and Multimodal Corpus Analysis in Audio Description: TRACCE'. In Remael, A., Orero, P. and Carroll, M. (eds.) *Audiovisual Translation and Media Accessibility at the Crossroads*. Amsterdam: Rodopi, pp. 409-421.
- Kazai, G., Kamps, J. and Milic-Frayling, N. (2012) *The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy*. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM'12, October 29th –November 2nd, Maui, Hawaii, USA.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D., Bernstein, M.S., Fei-Fei, L. (2017) 'Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations'. *International Journal of Computer Vision*, pp. 32–73.
- Kruger, J. L. (2010) 'Audio Narration: Re-Narrativising Film'. *Perspectives: Studies in Translatology*, 18, pp. 231-249.
- Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., & Luo, J. (2016). *TGIF: A new dataset and benchmark on animated GIF description*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 2016-December, pp. 4641–4650). IEEE Computer Society.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2015) *Microsoft COCO: Common Objects in Context*. Computer Vision, ECCV 2014, pp. 740–755.
- Little Miss Sunshine (2006) Directed by Jonathon Dayton and Valerie Faris [Film]. USA: 20th Century Fox Home Entertainment.
- Matamala, A. (2019) 'The ADLAB PRO Course Materials: Structure, Type, Quantity and Aims'. Multiplier Event 6, 3rd June, Ljubljana. URL: <https://www.adlabpro.eu/results/multiplier-events/multiplier-event-6/>.
- Pretty Woman (1990) Directed by Garry Marshall [Film]. USA: Touchstone Pictures.
- Ramos Caro, M. (2016) 'Testing Audio Narration: the Emotional Impact of Language in Audio Description'. *Perspectives, Studies in Translatology*, 24 (4), pp. 606-634.
- Rohrbach, A., Rohrbach, M., Tandon, N. and Schiele, B. (2015) 'A Dataset for Movie Description'. In *CVPR 2015*. URL: <https://www.cv-foundation.org/openaccess/CVPR2015.py>
- Ronchi, M. and Perona, P. (2015) *Describing Common Human Visual Actions in Images*. URL: [arXiv:1506.02203v1](https://arxiv.org/abs/1506.02203v1) [cs.CV].
- Salway, A. (2007) 'A Corpus-based Analysis of Audio Description', in Díaz-Cintas, J., Orero, P. and Remael, A. (eds.) *Media for All: Subtitling for the Deaf, Audio Description and Sign Language*. Amsterdam: Rodopi, pp. 151-174.
- Sjöberg, M., Tavakoli, H. R., Xu, Z., Mantecón, H. L., and Laaksonen, J. (2018) PicSOM Experiments in TRECVID 2018. In *Proceedings of the TRECVID 2018 Workshop*, Gaithersburg, USA.
- Sketch Engine (undated) 'Keywords and term extraction – identifying typical words'. URL: <https://www.sketchengine.eu/guide/keywords-and-term-extraction/?highlight=keywords>.

- Vandaele, J. (2012) 'What Meets the Eye. Cognitive Narratology for Audio Description'. *Perspectives: Studies in Translatology*, 20 (1), pp. 87-102.
- Vercauteren, G. (2007) 'Towards a European Guideline for Audio Description', in Díaz-Cintas, J., Orero, P. and Remael, A. (eds.) *Media for All: Subtitling for the Deaf, Audio Description and Sign Language*. Amsterdam: Rodopi, pp. 139-149.
- Yeung, J. (2007) 'Audio description in the Chinese world' in Díaz Cintas, J., Orero, P. and Remael, A. (eds.) *Media for All*. Amsterdam: Rodopi, pp. 231-244.