

Finding the Right Words: Investigating Machine-Generated Video Description Quality Using a Corpus-Based Approach

 Sabine Braun 

University of Surrey

 Kim Starr 

University of Surrey

Abstract

This paper examines first steps in identifying and compiling human-generated corpora for the purpose of determining the quality of computer-generated video descriptions. This is part of a study whose general ambition is to broaden the reach of accessible audiovisual content through semi-automation of its description for the benefit of both end-users (content consumers) and industry professionals (content creators). Working in parallel with machine-derived video and image description datasets created for the purposes of advancing computer vision research, such as Microsoft COCO (Lin et al., 2015) and TGIF (Li et al., 2016), we examine the usefulness of audio descriptive texts as a direct comparator. Cognisant of the limitations of this approach, we also explore alternative human-generated video description datasets including bespoke content description. Our research forms part of the MeMAD (Methods for Managing Audiovisual Data) project, funded by the EU *Horizon 2020* programme.

Key words: computer vision, machine learning, accessibility, audiovisual content, audio description, content description, content retrieval, video description, audiovisual translation, MeMAD.

Citation: Braun, S. & Starr, K. (2019). Finding the right words: Investigating machine-generated video description quality using a corpus-based approach. *Journal of Audiovisual Translation*, 2(2), 11–35.

Editor(s): G.M. Greco & A. Jankowska

Received: October 01, 2019


Accepted: November 19, 2019

Published: December 31, 2019

Funding: This publication is part of the EU funded project MeMAD, grant agreement number 780069.

Copyright: ©2019 Braun & Starr. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/). This allows for unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

 s.braun@surrey.ac.uk, <https://orcid.org/0000-0002-6187-3812>

 k.starr@surrey.ac.uk, <https://orcid.org/0000-0001-5236-1535>

1. Background

Audio description (AD) has established itself as a media service which facilitates access to audiovisual content for visually impaired audiences. Relying heavily on human resource, AD is currently an expensive part of the post-production process for traditional media companies, making it challenging to provide comprehensive media access (Sade, Naz, & Plaza, 2012, p. 270). The recent increase in user-generated audiovisual content has created a further challenge for media access. In parallel, research on automating the description of still images (“image captioning”) and video scenes (“video captioning”) has intensified and has begun to show moderate success (Krishna et al., 2017; Aafaq et al., 2019). The question of whether and to what extent these automated methods of description can be drawn upon to produce AD in order to reduce costs, and broaden media access without compromising quality, is therefore an economically and socially important question for research (Rohrbach et al., 2015b, p. 1).

Equally as important in this debate is the contribution to be derived from human video description analysis which has the potential to propel computer vision beyond standard object-and-action recognition tasks into the realm of multi-character, sequentially relayed narrative. However, since this is uncharted territory, new methods are required to bridge the human-computer void, bringing together the scientific world of algorithms and feature-extraction models with a humanities approach to cognition and understanding through a typically human lens.

The first step along this road is to form a deeper awareness of the distance that exists between human and machine, starting with a study of the main characteristics of automated image and video captions. Identifying patterns and behaviours that appear atypical in the context of human understanding allow us to isolate those areas of computer vision development which require most attention.

The current quality of image and video captioning and the absence of principles guiding the development of such captions raises ethical and quality issues from the customer perspective, as well as questions about the relevance and value of the “human voice” in this process. The “human” role is one aspect that the H2020 Project MeMAD (Methods for Managing Audiovisual Data: Combining Machine Efficiency and Human Accuracy)¹ is currently investigating. Our primary research focus in this project is to explore how our knowledge of human approaches to relevance and saliency in information selection can be used for modelling and improving the automation of video captioning in the fields of: (i) archive media retrieval; and (ii) AD for audiovisual media consumers. Whilst these practices overlap to some extent, the main driver for producing content descriptions for archival purposes is the likelihood of the re-use of the content internally or for re-sale to other media companies. Content descriptions for archival purposes therefore tend to be more “literal” or factual

¹ 2018–2020, Grant number 780069.

than AD, especially AD for filmic drama and movies, which is often “narrative” or figurative (Kruger, 2010; Ramos Caro, 2016). Whilst the (semi-)automation of content description is therefore likely to be a more achievable goal in the shorter term than a model for generating elaborate audio descriptions, the MeMAD project has adopted a two-pronged approach exploring content description for retrieval guided by work with broadcast archive journalists, as well as exploring how human knowledge can be used to support the (semi-)automation of video captioning in the context of enhancing traditional AD user experience. This paper focuses on the latter, while mindful of the former.

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. 2015a; 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 performs the task 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 (Audetel/ITC, 2000; AENOR, 2005), as there is a lack of consensus between describers about what should be included and omitted (Vercauteren, 2007, p. 139; Yeung, 2007, p. 241; Ibanez, 2010, p. 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 (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 machines are currently capable of producing (non-interpretive, observational, object/action oriented).

This paper outlines our approach to selecting and compiling appropriate corpora for this purpose and reports the outcomes of an initial comparison of human and machine-generated descriptions with regard to the quality of the descriptions. The final section will then discuss our findings and draw attention to the social and ethical implications that arise from these findings with regard to the automation of audiovisual content descriptions in the context of media accessibility.

2. 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 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.

2.1. Creating the MeMAD 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 hypothesised, 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 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:

Table 1.

Story Arc Parameters

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 seconds – 3 minutes	Limited duration story arcs should simplify sequence modelling
Storyline	Self-contained micro-narrative	e.g., initiating action/crisis, proposed solution, action based on solution, consequence, result
Objects	Unlimited	Although no limitation was put on the number of objects in an extract, only those objects regarded as key to the action were included in our annotations

A sample story arc, *Boy in a Field*, taken from the film *Little Miss Sunshine*, is illustrated in Figure 1. At the beginning of the extract a dispute arises between a teenage boy and his family. The dispute is subsequently resolved by the intervention of a young female family member. Screenshots of narratively key frames from the scene sit alongside a brief description of the action, provided in linear fashion:

Figure 1.

Sample Story Arc: Boy in a Field (Little Miss Sunshine)



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



Duane's young sister (Olive) offers to talk to him. She leaves the rest of the family back at the roadside and walks down a grassy slope towards her brother.



Olive crouches down behind Duane, and without speaking ...



... **puts** an arm around him, leaning her head tenderly on his shoulder.



Comforted by her presence and the knowledge that she truly understands his despair, Duane relinquishes his anger. They both rise ...



... **and** walk back towards the roadside where the rest of the family are waiting for them.



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

2.2. Audio Description

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 via AD is reliable and considered (i.e. contains minimal errors either in comprehension of source materials or exposition in the AD output).

2.3. Content Descriptions

Having established 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, metadata and ancillary texts (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 our study aimed at enhancing automated description services, the creation of a content descriptions corpus from 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 generate 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 key to 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 “ground truth” against which machine descriptions, governed by similar limitations inherent within the automation model, might be critically evaluated. Predictably, however, lexical variation within the AD is 29.66% greater when measured against the CD corpus (word-types) reflecting the more filmic, descriptive remit prevailing in most AD guidelines.

2.4. Training Data and Production of Captions for the MVC

A first-iteration corpus of captions (machine descriptions) was created by applying the MeMAD *DeepCaption* model (Sjöberg, Tavakoli, Xu, Mantecón, and Laaksonen, 2018), trained on image recognition using two large open access datasets, MS COCO (Lin et al., 2015) and TGIF (Li et al., 2016), to the MeMAD Video Corpus (MVC). 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 entirely 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.

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 separate operatives (Chen et al., 2015), as shown in

Figure 2. TGIF consists of 100k short sequence animated images (GIFs) drawn from *Tumblr* and annotated with 120k natural language sentences. Both MSCOCO and TGIF harnessed the power of crowdsourcing to produce the annotations.

Figure 2.

Example of captioned image from MS COCO



(338317)

- i. There is a lot of foot traffic on this street during the day.
- ii. People walking down a sidewalk near a road and a building.
- iii. A street with various people walking by a building.
- iv. There are people that are walking on the street
- v. An image of a person walking down the street on her phone

2.5. Annotation Procedure








With regard to the methodological approach to the creation of CD, our “story-arc” annotators were drawn from a pool of doctoral students and post-doctoral researchers experienced in multimedia research. Each annotated extract was verified for accuracy by an alternate annotator to the one creating the original file. Further operatives ensured standardisation of annotations in terms of lexicon and terminology, and performed text normalisation and data cleansing tasks. Each film extract and the associated annotations were therefore checked by three independent operatives before being admitted to the final MeMAD Video Corpus.

The AD and the dialogues were transcribed from the original film tracks. The video captions were produced in electronic text format by our project partner, Aalto University Computing Department. The three corpora were aligned at clip level to allow for direct comparison of the different types of description/annotation at this level. Further detail about the corpus creation, annotation and alignment is given in Braun, Starr and Laaksonen (2020).

Using the story arc introduced earlier, Figure 3 shows an example of the descriptions/annotations. Corpus analysis software *SketchEngine* (Kilgarriff et al., 2014) was used to compute basic descriptive statistics, which will be presented in the next section.

Figure 3.

Sample MVC Annotation

Frame/Time codes	Audio Description (AD) / Dialogue	Content Description (CD)	Machine Description (MD)
02:100994/01:07:19.760 		Dwayne is sitting on the grass in a field, hugging his knees. He is sitting with his back to us.	a man is sitting in a field
02:101125/01:07:25 	He is sitting with his back to her, arms resting on his knees, gazing at the rocky soil at his feet, and doesn't turn as she comes near.	Olive walks towards Dwayne, who is sitting on the ground, staring at the grass. Sheryl, Frank and Richard are at the top of the slope, standing next to the van, looking down at them.	a man and a woman are talking to each other
02:101650/01:07:46.000 	Dressed in her red T-shirt, pink shorts and red cowboy boots, her long hair tied back, her huge glasses perched on her nose, Olive squats at Dwayne's side.	Once she has reached Dwayne, Olive slows down and bends her knees to sit next to Dwayne. Dwayne does not react.	a group of people are singing and dancing
02:101875/01:07:55.000 	She puts her arm around him and rests her head on his shoulder. His head turns slightly towards her.	Olive looks at Dwayne and then puts her arm around him, resting her head on his shoulder. Dwayne is trying not to cry.	a group of people are in a field
02:102325/01:08:13.000 	Dwayne: I'm OK... let's go.	Dwayne turns towards Olive. Dwayne reassures Olive that he is okay, and she looks at him and smiles.	a man is running
02:102475/01:08:19.000 	Olive stands up and Dwayne gets to his feet and goes with her to the bottom of the slope.	Olive and Dwayne stand up and slowly walk towards the bottom of the slope.	a man and a woman are walking in a field
02:102625/01:08:25.000 	Olive starts to climb, putting out her hand for support. Dwayne lifts her up underneath her arms and carries her to the top of the slope.	Olive climbs the slope but she wobbles. Dwayne helps her by carrying her up. Olive seems to be smiling.	a woman is walking down the road

2.6. Initial Corpus Comparison

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 arbitrary and could easily be changed if the frequency/points at which the machine produces a caption is adjusted. As explained above, a caption was generated for the middle of each shot.

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: 580; CD: 3,061; AD: 3,969), 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 0.044.²

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 (MD). 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

² Due to the much larger size of the TIWO (over 300k words), this is only a rough indicator, as it is natural for the TTR to decrease with corpus size. However, the different genres may have had an impact.

might potentially be partially rectified in future machine iterations by changes to computer vision feature extraction.

Table 2.

Corpus Comparison

<i>Category</i>	<i>MD</i>	<i>MD</i>	<i>CD</i>	<i>CD</i>	<i>AD</i>	<i>AD</i>
	<i>Types</i>	<i>Tokens</i>	<i>Types</i>	<i>Tokens</i>	<i>Types</i>	<i>Tokens</i>
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

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 the assessment of the current quality of machine-generated descriptions.

3. Video Captions: Quality Assessment

Our initial quantitative analyses of the machine-generated descriptions, as exemplified in Table 2, 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.

3.1. 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, Filatova, and Ipeirotis, 2018). Research shows that the workers' profile has an impact on the quality of their work (Kazai, Kamps, and Milic-Frayling 2012) and that feedback can improve quality (Han, Roitero, & Gadiraju 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 4 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, not giving much detail about the actual 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.

Figure 4.

Examples of Captioned Images from MS COCO

	<p>1 (374628)</p> <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.
	<p>2 (132394)</p> <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
	<p>3 (290868)</p> <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

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 2 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 the 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 4, 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 after dinner, 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 arguing”, although even then the saliency may not relate either to the household chore, or the argument, but instead 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. 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”.

Interestingly, while AD may assist in determining that a man and a woman are in the kitchen (the fact that they are arguing 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, kitchen, washing up, angry faces, aggressive body language, arguing – 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.

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 5. The story arc from which it is taken shows one man and one woman.

Figure 5.

Example of Machine Description from MVC Clip 200006

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

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.

3.2. 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 a significant error rate. 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).

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. Equally curious, changes in the pixel structure of an image that cannot be detected by the human eye can change the descriptions in an unpredictable fashion.

Unusual or rare objects pose a further challenge as do facial expressions: laughing and grinning are difficult to distinguish in current models. More training data is needed to overcome these difficulties although, again, audio cues could assist once incorporated into the model.

3.3. 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:

Table 3.

MD Corpus: Verb Rankings

MD Corpus Verb Rank	Lemma	Frequency	MD Corpus Verb Rank	Lemma	Frequency
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

Eighty-eight verb lemmas can be found in the MD lexicon, only forty-six of which occur five or more times (see Table 3). 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 (Table 4.), suggesting that feature extraction and other factors play a significant role.

Table 4.

MD Verbs: Comparative Statistics vs. Training Datasets

RANK	VERB LEMMA	MD <i>f</i>	MD verb/m	COCO <i>f</i>	COCO verb/m	TGIF <i>f</i>	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

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. 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.

Table 5.

MD Corpus: Keyness Scores

Rank	Term	Score	(MD) corpus frequency	Reference (BNC) corpus 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

4. Conclusions: Quality Issues and the Future of AI for AD

The ongoing AI revolution has the potential to promote inclusive design, by personalising media products (James, 2019) and making them accessible for everyone, bridging language barriers as well as different physical and cognitive abilities. In the context of audiovisual accessibility, this is, however,

a long way off. 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. This article has highlighted problems with object recognition, gender labelling, action interpretation and so forth. However, saliency, relevance and lack of narrative coherence emerge as fundamental issues (Huang et al., 2016). Currently, MDs not only fail to approach human levels of description, in terms of complex artificial cognition such as mental modelling, but they also fall at the first hurdle (e.g., mistaking a desk for a surfboard). Resolving basic computer vision (objects and actions) will therefore only solve some of the more immediate problems associated with complex narrative.

As was shown, the quality of MDs is currently affected by a lack of sufficient training data, especially moving imagery (Aafaq et al., 2019), and this further impacts the inter-relational linking of artefacts for narrative development, facial recognition, facial expression and emotion detection, amongst many other factors. One important finding of this exploratory research is reinforcement of the need for further relevant training datasets to be created, despite their limitations. What the present article has highlighted in this respect is that AD is not directly comparable with MD, and alternative human-derived datasets are more helpful for training the model. As discussed, CD appears to be 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 most important 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 (Ofcom, 2017, Annex 4, p. 18). However, lower quality 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.

Human approaches to audiovisual content description will continue to drive the agenda and establish priorities for future computer vision research. Semi-automation and post-editing afford further opportunities to enhance the machine's best efforts, with human-in-the-loop approaches being used to determine how human and machine intelligence can most productively and efficiently come together. Combining human and machine endeavours will also demonstrate to the human creators of AD that their involvement in developing semi-automated approaches will not mean that they are writing themselves out of their jobs, and indeed, is more likely to secure their involvement in the development of automated approaches where these can be useful. For instance, automation or semi-automation of AD carries enormous potential in the area of social media (YouTube; Facebook; Twitter images/gifs; Instagram) and in other multimodal information situations e.g., language learning and pedagogy more generally. The addition of AD to a range of social media content would serve the purpose of making AD more widely available generally.

This could be particularly beneficial if machines delivered AD where there is currently no alternative (low quality arguably being better than zero access). Taking this path will contribute to improving media accessibility for everyone while simultaneously invoking reflective practices and a mindful approach to the social, ethical and economic implications of automation in this area.

References

- 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. Retrieved from: <https://www.aenor.com/normas-y-libros/buscador-de-normas/UNE?c=N0032787>
- Aafaq, N., Milan, A., Liu, W., Gilani, S.Z., and Shah, M. (2019). *Video description: A survey of methods, datasets and evaluation metrics* (ACM Computing Surveys). Retrieved from: [arXiv:1806.00186v1](https://arxiv.org/abs/1806.00186v1) [cs.CV].
- Braun, S. (2011). Creating coherence in audio description. *Meta*, 56(3), 645–662.
- Braun, S. (2016). The importance of being relevant? A cognitive-pragmatic framework for conceptualising audiovisual translation. *Target*, 28(2), 302–313.
- Braun, S., Starr, K., and Laaksonen, J. (2020, forthcoming). Comparing human and automated approaches to visual storytelling. In S. Braun, and 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. Retrieved from: [arXiv:1504.00325v2](https://arxiv.org/abs/1504.00325v2) [cs.CV].
- Difallah, D., Filatova, E., and Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, February 5–9th. Marina Del Rey, CA: ACM.
- Fresno, N., Castellà, J., and Soler-Vilageliu, O. (2016). What should I say? Tentative criteria to prioritize information in the audio description of film characters. In A. Matamala and P. Orero, (Eds.) *Researching audio description* (pp. 143–167). London: Palgrave.
- Han, L., Roitero, K., and Gadiraju, U. (2019). All those wasted hours: On task abandonment in crowdsourcing. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 321–329), 11–15th February. Melbourne, Australia: ACM.
- 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. *Proceedings of the NAACL-HLT*, June 12th–17th. San Diego, CA: Association for Computational Linguistics.
- Hyks, V. (2005). Audio description and translation: Two related but different skills. *Translating Today*, 4, 6–8.
- Ibanez, A. (2010). Evaluation criteria and film narrative. A frame to teaching relevance in audio description. *Perspectives: Studies in Translatology*, 18(3), 143–153.
- Independent Television Commission (2000). *Guidance on standards for audio description*. Retrieved from: https://www.audiodescription.co.uk/uploads/general/itcguide_sds_audio_desc_word3.pdf.

- James, L. (2019, 12 August). Getting personal with artificial intelligence and the cloud. *The Record*. Retrieved from: <https://www.technologyrecord.com/Article/getting-personal-with-artificial-intelligence-and-the-cloud-85390>.
- Jimenez, C., and Seibel, C. (2012). Multisemiotic and multimodal corpus analysis in audio description: TRACCE. In A., Remael, P. Orero, and M. Carroll (Eds.) *Audiovisual translation and media accessibility at the crossroads* (pp. 409–421). Amsterdam: Rodopi.
- Kazai, G., Kamps, J., and Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM'12*, October 29 – November 2nd. Maui, HI: ACM.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, (1), 7–36.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D., Bernstein, M. S., and Fei-Fei, L. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 32–73.
- Kruger, J. L. (2010). Audio narration: Re-narrativising film. *Perspectives: Studies in Translatology*, 18, 231–249.
- Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., and Luo, J. (2016). TGIF: A new dataset and benchmark on animated GIF description. *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (Vol. 2016 – December, 4641–4650). Las Vegas, NV: IEEE Computer Society.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girschick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common objects in context. *Computer Vision, ECCV 2014*, 740–755.
- Matamala, A. (2019). The ADLAB PRO course materials: Structure, type, quantity and aims. Multipler event 6, Ljubljana, 3rd June. Retrieved from: <https://www.adlabpro.eu/results/multiplier-events/multiplier-event-6/>
- Ofcom (2017). Ofcom's code on television access services. Retrieved from: <https://www.ofcom.org.uk/tv-radio-and-on-demand/broadcast-codes/tv-access-services>
- Ramos Caro, M. (2016). Testing audio narration: The emotional impact of language in audio description. *Perspectives: Studies in Translatology*, 24(4), 606–634.
- Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (2015a). A dataset for movie description. *Proceedings of CVPR*, Boston, Mass, June 8–10th. Retrieved from: <https://www.cv-foundation.org/openaccess/CVPR2015.py>
- Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., and Schiele, B. (2015b). Movie description. *Proceedings of ICCV*, December 11–18th. Santiago, Chile. Retrieved from: [arXiv:1605.03705v1](https://arxiv.org/abs/1605.03705v1) [cs.CV].
- Ronchi, M. and Perona, P. (2015). *Describing common human visual actions in images*. Retrieved from: [arXiv:1506.02203v1](https://arxiv.org/abs/1506.02203v1) [cs.CV].
- Sade, J., Naz, K., and Plaza, M. (2012). Enhancing audio description: A value added approach. *Proceedings of ICCHP, Part 1, LNCS 7382* (pp. 270–277). Linz, Austria: ICCHP.
- Salway, A. (2007). A corpus-based analysis of audio description. In J. Díaz-Cintas, P. Orero, and A. Remael (Eds.) *Media for all: Subtitling for the deaf, audio description and sign language* (pp. 151–174). Amsterdam: Rodopi.

- Sjöberg, M., Tavakoli, H. R., Xu, Z., Mantecón, H. L., and Laaksonen, J. (2018). PicSOM Experiments in TRECVID 2018. *Proceedings of the TRECVID 2018 Workshop*. Gaithersburg, MD.
- Vandaele, J. (2012). What meets the eye. Cognitive narratology for audio description. *Perspectives: Studies in Translatology*, 20(1), 87–102.
- Vercauteren, G. (2007). Towards a European guideline for audio description. In J. Díaz-Cintas, P. Orero, and A. Remael (Eds.) *Media for all: Subtitling for the deaf, audio description and sign language* (pp. 139–149). Amsterdam: Rodopi.
- Yeung, J. (2007). Audio description in the Chinese world. In J. Díaz Cintas, P. Orero, and A. Remael (Eds.) *Media for all: Subtitling for the deaf, audio description and sign language* (pp. 231–244). Amsterdam: Rodopi.